日　本　国　特　許　庁
JAPAN　PATENT　OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office

出願年月日
Date of Application:　２００３年　２月２７日

出願番号
Application Number:　特願２００３－０５０２４４

[ST.10/C]:　[ＪＰ２００３－０５０２４４]

出願人
Applicant(s):　株式会社日立製作所

２００３年　４月２５日

特許庁長官
Commissioner,
Japan Patent Office　太田信一郎

出証番号　出証特２００３－３０３０７３７

【書類名】　特許願

【整理番号】　K02017501A

【あて先】　特許庁長官殿

【国際特許分類】　G06F　3/06

【発明者】

【住所又は居所】　神奈川県小田原市中里３２２番地２号　株式会社日立製作所　ＲＡＩＤシステム事業部内

【氏名】　武田　貴彦

【発明者】

【住所又は居所】　神奈川県小田原市中里３２２番地２号　株式会社日立製作所　ＲＡＩＤシステム事業部内

【氏名】　安積　義弘

【発明者】

【住所又は居所】　神奈川県小田原市中里３２２番地２号　株式会社日立製作所　ＲＡＩＤシステム事業部内

【氏名】　山神　憲司

【発明者】

【住所又は居所】　神奈川県小田原市中里３２２番地２号　株式会社日立製作所　ＲＡＩＤシステム事業部内

【氏名】　鈴木　勝喜

【発明者】

【住所又は居所】　神奈川県川崎市麻生区王禅寺１０９９番地　株式会社日立製作所　システム開発研究所内

【氏名】　白銀　哲也

【特許出願人】

【識別番号】　000005108

【氏名又は名称】　株式会社日立製作所

【代理人】

【識別番号】　100075096

【弁理士】

【氏名又は名称】　作田　康夫

【手数料の表示】

【予納台帳番号】　013088

【納付金額】　21,000円

【提出物件の目録】

【物件名】　明細書　1

【物件名】　図面　1

【物件名】　要約書　1

【プルーフの要否】　要

【書類名】　明細書

【発明の名称】　データ処理システム

【特許請求の範囲】

【請求項１】

第一の計算機及び前記第一の計算機と接続される第一のストレージサブシステムを有する第一のストレージシステムと、

第二の計算機及び前記第二の計算機と接続される第二のストレージサブシステムを有する第二のストレージシステムとを有し、

前記第一及び第二の計算機は第一の通信線で接続され、

前記第一のストレージサブシステムと前記第二のストレージサブシステムは第二の通信線で接続されており、

前記第一のストレージサブシステムは、前記第一の計算機の指示に従って、前記第一のストレージサブシステムが有する第一の記憶領域に対するデータの更新をジャーナルとして第二の記憶領域に記録し、

前記第一の計算機は、前記第一の通信線を介して前記第二の計算機に前記第一のストレージサブシステムにおけるジャーナルの蓄積に関する情報を送信し、

前記第二の計算機は、受信した前記ジャーナルの蓄積に関する情報に基づいて、前記第二のストレージサブシステムに対するコマンドを発行し、

前記第二のストレージサブシステムは、前記コマンドに応答して、前記第一のストレージサブシステムに対して、前記ジャーナルの送付を要求するコマンドを送信し、

前記第一のストレージサブシステムは、前記第二のストレージサブシステムからのコマンドに応答して、要求されたジャーナルを前記第二の通信線を介して転送し、

前記第二のストレージサブシステムは、前記転送されたジャーナルを該第二のストレージサブシステムが有する第三の記憶領域に格納することを特徴とするデータ処理システム。

【請求項２】

請求項１記載のデータ処理システムであって、

前記第二のストレージサブシステムは、前記第一のストレージサブシステムの前記第一の記憶領域に対応する第四の記憶領域を有し、

前記第二のストレージサブシステムは、前記第二の計算機の指示に従って、前記第一の記憶領域に格納されているデータの転送を要求する第二のコマンドを前記第一のストレージサブシステムに送信し、

前記第二のコマンドに従って送信されたデータを前記第四の記憶領域に格納することを特徴とするデータ処理システム。

【請求項３】

請求項２記載のデータ処理システムであって、

前記第二のストレージサブシステムは、前記第二の計算機の指示に従って、前記第三の記憶領域に格納されたジャーナルを用いて前記第四の記憶領域に格納されたデータを更新することを特徴とするデータ処理システム。

【請求項４】

請求項１記載のデータ処理システムにおいて、

前記第二のストレージシステムが発行する前記第二のストレージサブシステムに対するコマンドには、

前記第一のストレージサブシステムを指定する情報、前記第二のストレージサブシステムを指定する情報、前記第二及び第三の記憶領域を指定する情報、前記第二及び第三の記憶領域の部分を指定する情報が含まれていることを特徴とするデータ処理システム。

【請求項５】

請求項１記載のデータ処理システムにおいて、前記第二の計算機が前記第二のストレージサブシステムへ発行する前記コマンドの発行のタイミングは、ユーザによって指定される任意のタイミングであることを特徴とするデータ処理システム。

【請求項６】

請求項３記載のデータ処理システムにおいて、

前記第二の計算機の指示は、所定のタイミングで行われることを特徴とするデータ処理システム。

【請求項７】

請求項３記載のデータ処理システムにおいて、

前記ジャーナルには、前記第一の記憶領域を更新した時間に関する情報が含まれ、

前記第二の計算機の指示には、所定の時間に関する情報が含まれ、

前記第二のストレージサブシステムは、前記所定の時間以前に更新された前記第一の記憶領域に関するジャーナルを用いて、前記第四の記憶領域に格納されたデータを更新することを特徴とするデータ処理システム。

【請求項８】

請求項２記載のデータ処理システムにおいて、

前記第一、第二、第三又は第四の記憶領域が、複数の論理ボリュームから構成されていることを特徴とするデータ処理システム。

【請求項９】

請求項２記載のデータ処理システムにおいて、

前記第二の記憶領域は、前記第一の計算機で実行されるユーザアプリケーションプログラムには使用されない記憶領域であることを特徴とするデータ処理システム。

【請求項１０】

請求項１記載のデータ処理システムにおいて、前記第二のストレージサブシステムは、前記第一のストレージサブシステムの前記第一の記憶領域に対応する第四の記憶領域を有し、

前記第二のストレージシステムは、前記第三の記憶領域に格納されたジャーナルを読み出し、読み出したジャーナルを用いて、前記第四の記憶領域に格納されたデータを更新することを特徴とするデータ処理システム。

【請求項１１】

第一の計算機及び前記第一の計算機と接続される第一のストレージサブシステムを有する第一のストレージシステムと、

第二の計算機及び前記第二の計算機と接続される第二のストレージサブシステムを有する第二のストレージシステムとを含むデータ処理システムにおいて、

前記第一及び第二の計算機は第一の通信線で接続され、

前記第一のストレージサブシステムと前記第二のストレージサブシステムは第二の通信線で接続されており、

前記第一のストレージサブシステムは、前記第一の計算機の指示に従って、前記第一のストレージサブシステムが有する第一の記憶領域に対するデータの更新をジャーナルとして第二の記憶領域に記録し、

前記第一の計算機は、前記ジャーナルの蓄積に関する情報を前記第一のストレージサブシステムから取得し、

前記第一の計算機は、前記取得された前記ジャーナルの蓄積に関する情報に基づいて、前記第一のストレージサブシステムに対するコマンドを発行し、

前記第一のストレージサブシステムは、前記コマンドに応答して、前記第二のストレージサブシステムに対して、前記ジャーナルを送信し、

前記第二のストレージサブシステムは、前記第二の通信線を介して転送されたジャーナルを該第二のストレージサブシステムが有する第三の記憶領域に格納することを特徴とするデータ処理システム。

【請求項１２】

請求項１１記載のデータ処理システムにおいて、前記第二のストレージサブシステムは、前記第一のストレージサブシステムの前記第一の記憶領域に対応する第四の記憶領域を有し、

前記第二のストレージシステムは、前記第三の記憶領域に格納されたジャーナルを読み出し、読み出したジャーナルを用いて、前記第四の記憶領域に格納されたデータを更新することを特徴とするデータ処理システム。

【請求項１３】

請求項１記載のデータ処理システムにおいて、

さらに第三の計算機及び前記第三の計算機と接続される第三のストレージサブシステムを有する第三のストレージシステムを有し、

前記第一の計算機と前記第三の計算機は前記第一の通信線を介して接続されており、

前記第一のストレージサブシステムと前記第三のストレージサブシステムは第

三の通信線で接続されており、

前記第一の計算機は、前記ジャーナルの蓄積に関する情報を前記第一の通信線を介して前記第三の計算機に送信し、

前記第三の計算機は、受信した前記ジャーナルの蓄積に関する情報に基づいて、前記第三のストレージサブシステムに対するコマンドを発行し、

前記第三のストレージサブシステムは、前記コマンドに応答して、前記第一の記憶装置システムに対して、前記ジャーナルの送付を要求するコマンドを送信し、

前記第一のストレージサブシステムは、前記第三のストレージサブシステムからのコマンドに応答して、要求されたジャーナルを前記第三の通信線を介して転送し、

前記第三のストレージサブシステムは、前記転送されたジャーナルを該第三のストレージサブシステムが有する第五の記憶領域に格納することを特徴とするデータ処理システム。

【請求項１４】

第一の計算機及び前記第一の計算機と接続される第一のストレージサブシステムを有する第一のストレージシステムと、

第二の計算機及び前記第二の計算機と接続される第二のストレージサブシステムを有する第二のストレージシステムと、

前記第一の計算機システム及び前記第二の計算機システムに接続される記憶装置とを有し、

前記第一及び第二の計算機は通信線で接続されており、

前記第一のストレージサブシステムは、前記第一の計算機の指示に従って、前記第一のストレージサブシステムが有する第一の記憶領域に対するデータの更新をジャーナルとして第二の記憶領域及び前記記憶装置に記録し、

前記第一の計算機は、前記通信線を介して前記第二の計算機に前記第一のストレージサブシステムにおけるジャーナルの蓄積に関する情報を送信し、

前記第二の計算機は、受信した前記ジャーナルの蓄積に関する情報に基づいて、前記第二のストレージサブシステムに対するコマンドを発行し、

前記第二のストレージサブシステムは、前記コマンドに応答して、前記記憶装置から前記ジャーナルを読み出し、

前記第二のストレージサブシステムは、前記転送されたジャーナルを該第二のストレージサブシステムが有する第三の記憶領域に格納することを特徴とするデータ処理システム。

【請求項１５】

請求項１４記載のデータ処理システムであって、

前記第二のストレージサブシステムは、前記第一のストレージサブシステムの前記第一の記憶領域に対応する第四の記憶領域を有し、

前記第一のストレージサブシステムは、前記第一の計算機の指示に従って、前記第一の記憶領域に格納されているデータを前記記憶装置へ転送し、

前記第二のストレージサブシステムは、前記第二の計算機の指示に従って、前記第一の記憶領域に格納されているデータの転送を要求する第二のコマンドを前記記憶装置に送信し、

前記第二のコマンドに従って送信されたデータを前記第四の記憶領域に格納することを特徴とするデータ処理システム。

【請求項１６】

請求項１５記載のデータ処理システムであって、

前記第二のストレージサブシステムは、前記第二の計算機の指示に従って、前記第三の記憶領域に格納されたジャーナルを用いて前記第四の記憶領域に格納されたデータを更新することを特徴とするデータ処理システム。

【請求項１７】

第一の計算機及び前記第一の計算機と接続される第一のストレージサブシステムを有する第一のストレージシステムと、

第二の計算機及び前記第二の計算機と接続される第二のストレージサブシステムを有する第二のストレージシステムとを有し、

前記第一及び第二の計算機は第一の通信線で接続され、

前記第一のストレージサブシステムは、さらに、第三の計算機及び第一の複数のストレージサブシステムを有し、

前記第三の計算機は、前記第一の複数のストレージサブシステムが有する記憶領域から第一及び第二の仮想的な記憶領域を作成して前記第一の計算機に提供し、

前記第二のストレージサブシステムは、さらに、第四の計算機及び第二の複数のストレージサブシステムを有し、

前記第四の計算機は、前記第二の複数のストレージサブシステムが有する記憶領域から第三及び第四の仮想的な記憶領域を作成して前記第二の計算機に提供し、

前記第三の計算機と前記第四の計算機は第二の通信線で接続されており、

前記第一のストレージサブシステムは、前記第一の計算機の指示に従って、前記第一の仮想的記憶領域に対するデータの更新をジャーナルとして前記第二の仮想的な記憶領域に記録し、

前記第一の計算機は、前記第一の通信線を介して前記第二の計算機に前記第一のストレージサブシステムにおけるジャーナルの蓄積に関する情報を送信し、

前記第二の計算機は、受信した前記ジャーナルの蓄積に関する情報に基づいて、前記第二のストレージサブシステムに対するコマンドを発行し、

前記第二のストレージサブシステムは、前記コマンドに応答して、前記第一のストレージサブシステムに対して、前記ジャーナルの送付を要求するコマンドを送信し、

前記第一のストレージサブシステムは、前記第二のストレージサブシステムからのコマンドに応答して、要求されたジャーナルを前記第二の通信線を介して転送し、

前記第二のストレージサブシステムは、前記転送されたジャーナルを前記第三の仮想的な記憶領域に格納することを特徴とするデータ処理システム。

【請求項１８】

請求項１記載のデータ処理システムにおいて、

さらに第三の計算機及び前記第三の計算機と接続される第三のストレージサブシステムを有する第三のストレージシステムを有し、

前記第一の計算機と前記第三の計算機は前記第一の通信線を介して接続され、

前記第一のストレージサブシステムと前記第三のストレージサブシステムは第三の通信線で接続されており、

前記第二の計算機又は前記第二のストレージサブシステムの障害を検出した前記第一の計算機が、前記第三の計算機に前記第一の通信線を介して前記ジャーナルの蓄積に関する情報を送信し、

前記第三の計算機は、受信した前記ジャーナルの蓄積に関する情報に基づいて、前記第三のストレージサブシステムに対するコマンドを発行し、

前記第三のストレージサブシステムは、前記コマンドに応答して、前記第一のストレージサブシステムに対して、前記ジャーナルの送付を要求するコマンドを送信し、

前記第一のストレージサブシステムは、前記第三のストレージサブシステムからのコマンドに応答して、要求されたジャーナルを前記第三の通信線を介して転送し、

前記第三のストレージサブシステムは、前記転送されたジャーナルを該第三のストレージサブシステムが有する第五の記憶領域に格納することを特徴とするデータ処理システム。

【請求項１９】

請求項１記載のデータ処理システムにおいて、

さらに第三の計算機及び前記第三の計算機と接続される第三のデータ処理システムを有する第三のストレージシステムを有し、

前記第二の計算機と前記第三の計算機は、前記第一の通信線を介して接続されており、

前記第一のストレージサブシステムと前記第三のストレージサブシステムは第三の通信線で接続されており、

前記第一の計算機又は前記第一のストレージサブシステムの障害を検出した前記第二の計算機が、前記第三の計算機に前記ジャーナルの蓄積に関する情報を前記第一の通信線を介して送信し、

前記第三の計算機は、受信した前記ジャーナルの蓄積に関する情報に基づいて、前記第三のストレージサブシステムに対するコマンドを発行し、

前記第三のストレージサブシステムは、前記コマンドに応答して、前記第二のストレージサブシステムに対して、前記ジャーナルの送付を要求するコマンドを送信し、

前記第二のストレージサブシステムは、前記第三のストレージサブシステムからのコマンドに応答して、要求されたジャーナルを前記第三の通信線を介して転送し、

前記第三のストレージサブシステムは、前記転送されたジャーナルを該第三のストレージサブシステムが有する第五の記憶領域に格納することを特徴とする計算機システム。

【請求項２０】

第一の計算機及び第一のストレージサブシステムを有するサービスプロバイダにおけるサービス提供方法であって、

ユーザが、前記第一の計算機を、前記ユーザが有する第二の計算機に登録し、

前記ユーザは、前記第二の計算機に接続される第三のストレージサブシステムに格納されたデータの複製を格納する第三のストレージサブシステムを有し、

前記第二の計算機は、前記第三のストレージサブシステム又は前記第三のストレージサブシステムに接続される第三の計算機の障害を検出した場合に前記第一の計算機を選択して前記第一の計算機及び前記第一のストレージサブシステムの使用を開始し、

前記サービスプロバイダは、前記第一の計算機及び前記第一のストレージサブシステムの使用の開始に伴って前記ユーザに課金することを特徴とする方法。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は、第一のストレージシステムに格納されたデータを、第二のストレージシステムに複製するための技術に関する。

【０００２】

【従来の技術】

近年、常に顧客に対して継続したサービスを提供するために、第一のストレー

ジシステムに障害が発生した場合でもデータ処理システムがサービスを提供できるよう、システム間でのデータの複製に関する技術が重要になっている。第一のストレージシステムに格納された情報を第二のストレージシステムに複製する技術としては、特許文献に開示された技術が存在する。特許文献には、第一のストレージシステムに含まれる計算機（以下、「プライマリホスト」という）が、プライマリホストに接続されたディスクアレイ装置（以下、「プライマリディスクアレイ装置」）に格納されたデータを、通信リンク及び第二のストレージシステムに含まれる計算機（以下、「セカンダリホスト」という）を介して、セカンダリホストに接続されたディスクアレイ装置（以下「セカンダリディスクアレイ装置」）に転送する技術（以下、「従来技術」という）が開示されている。

【０００３】

一方、コンピュータネットワークの発達により、企業等の有する情報処理システムがより複雑になっている。このことから、ネットワークに接続される機器を一括して管理したいという要求が高まっている。このような要求を満たす技術として、ファイバチャネルあるいはインターネット等のネットワークを介して接続された複数のストレージサブシステムを一括して管理し、仮想的に一つ又は複数のストレージサブシステムとしてユーザに提供するバーチャリゼーションという技術が考案されている。これにより、情報処理システムを使用するユーザは、複数のストレージサブシステムを、あたかも１つのストレージサブシステムであるかのように使用することができる。

【０００４】

以下、「ストレージサブシステム」は、単体のハードディスクドライブ、複数のハードディスクドライブの集合体、若しくは専用の制御部で複数のハードディスクドライブを制御するディスクアレイ装置等の記憶装置システムを示す。
以下、「ストレージシステム」及び「サイト」は、１つまたは複数のホストコンピュータと１つまたは複数のストレージサブシステムが接続された情報処理システムを示す。尚、以下、ホストコンピュータは単にホストと略す場合がある。

【０００５】

【特許文献１】

米国特許５１７０４８０号公報

【０００６】

【発明が解決しようとする課題】

ここで、上記従来技術を複雑化した情報処理システムに適用することを考える。

【０００７】

従来技術においては、プライマリホストとセカンダリホストの間で、双方のディスクアレイ装置に格納されたデータの転送が行われる。つまり、各ホストがデータ転送の経路として使用される。また、ネットワークに接続された各ホストは、ネットワークに接続される複数のディスクアレイ装置に関する情報（ネットワークアドレス等）を保持している。

【０００８】

したがって従来技術では、多数存在するディスクアレイ装置のいずれかを各ホストで適当に選択して、選択されたディスクアレイ装置にデータを容易に複製することが出来る。特に従来技術をバーチャリゼーションに適用する場合、バーチャリゼーションを管理している装置（計算機又はスイッチ）と各ホストとを連携させることができる。しかし、各ディスクアレイ装置に格納されたデータがホスト間の通信リンクを介して転送されるので、各ホストのチャネル負荷及びホスト間を接続する回線のトラフィックが増加するという問題点がある。

【０００９】

【課題を解決するための手段】

上記課題を解決するために、本発明では、プライマリ及びセカンダリホストが、各ホスト上で動作するソフトウエアに基づいて、各ホストに接続されたストレージサブシステム、例えば、各々プライマリ又はセカンダリディスクアレイ装置の状態を監視する。また、必要に応じて、プライマリ又はセカンダリホストは、ディスクアレイ装置間のデータ転送をプライマリ又はセカンダリディスクアレイ装置に指示する。さらに、各ホストは、各ディスクアレイ装置間のデータ転送を行うための情報をホスト間通信で遣り取りする。一方、各ディスクアレイ装置に格納されたデータは、ディスクアレイ装置間の専用線を介して転送される。尚、

専用線ではなく、テープ装置等の可搬記憶媒体を用いてデータを転送する形態も考えられる。

【００１０】

また、本発明においては、プライマリディスクアレイ装置は、プライマリディスクアレイ装置に格納されたデータの更新の情報をジャーナル（更新履歴）として格納する。ジャーナルは、具体的には、更新に用いられたデータのコピーとメタデータの記録である。さらに、プライマリディスクアレイ装置は、このジャーナルを、各ホストの指示にしたがって、セカンダリディスクアレイ装置に転送する。セカンダリディスクアレイ装置は、セカンダリホストの指示にしたがって、プライマリディスクアレイ装置から受取ったジャーナルを用いてセカンダリディスクアレイ装置に格納されたデータをプライマリディスクアレイ装置で更新されたのと同じ様に更新する。この更新は、プライマリディスクアレイ装置の更新をセカンダリディスクアレイ装置で再現することから、「リストア」と呼ばれる。

【００１１】

尚、ジャーナルの転送は、セカンダリディスクアレイ装置がプライマリディスクアレイ装置にジャーナルコピー命令を発行することで実現する構成としても良い。

【００１２】

また、各ホストに接続されるディスクアレイ装置は、仮想化を制御する装置によって仮想化されたストレージサブシステムでも良い。この場合、データ転送は、仮想化を制御する装置間で、あるいは、仮想化を制御する装置に接続されたストレージサブシステムと、他の同システムで行われる。

【００１３】

【発明の実施の形態】

図１は、本発明を適用したデータ処理システムの第一の実施形態のハードウェア構成を示す図である。

【００１４】

本データ処理システムは、プライマリホスト１００Ａ及びプライマリストレージサブシステム（プライマリディスクアレイ装置）２００Ａを有する第一のスト

レージシステム（以下「プライマリストレージシステム」または「プライマリサイト」）１０、セカンダリホスト１００Ｂ及びセカンダリストレージサブシステム（セカンダリディスクアレイ装置）２００Ｂを有する第二のストレージシステム（以下「セカンダリストレージシステム」または「セカンダリサイト」）２０、並びにリモートコンソール４０を有する。本実施形態では、ストレージサブシステムとしてディスクアレイ装置を例に説明するが、発明は特にディスクアレイ装置に限られない。プライマリサイトとプライマリディスクアレイ装置は各々第一のサイトと第一のディスクアレイ装置と呼ばれることがある。同様に、セカンダリサイトとセカンダリディスクアレイ装置は各々第二のサイトと第二のディスクアレイ装置と呼ばれることがある。

【００１５】

各ホスト１００（プライマリホスト１００Ａ及びセカンダリホスト１００Ｂ）は、ＣＰＵ１１０、主記憶装置１２０、及び入出力処理装置１３０を有する計算機である。具体的には、ワークステーション、マイクロコンピュータ又はメインフレームコンピュータ等である。

【００１６】

各ディスクアレイ装置２００（プライマリディスクアレイ装置２００Ａ及びセカンダリディスクアレイ装置２００Ｂ）は、記憶制御装置２１０、複数のディスク装置２２０及びＳＶＰ（Service Processor）２３０を有する。記憶制御装置２１０は、ホストアダプタ２１１、キャッシュメモリ２１２、ディスクアダプタ２１３、プロセッサ２１４、及び制御メモリ２１５を有する。

【００１７】

プライマリホスト１００Ａはプライマリディスクアレイ装置２００Ａと、セカンダリホスト１００Ｂはセカンダリディスクアレイ装置２００Ｂと、それぞれファイバチャネル６６によって接続される。各ホスト１００のＣＰＵ１１０および主記憶装置１２０は、入出力処理装置１３０及びファイバチャネル６６を介してディスクアレイ装置２００のホストアダプタ２１１に接続される。

【００１８】

また、プライマリディスクアレイ装置２００Ａとセカンダリディスクアレイ装

置２００Ｂは、ファイバチャネル６８を介して互いに接続される。プライマリディスクアレイ装置とセカンダリディスクアレイ装置は同じ部屋の中、または同じビルの中にあっても良いが、安全性のために（両装置が同時に同じ障害にあわないよう）距離を置いても良い。プライマリサイト１０とセカンダリサイト２０との間の距離が長い、すなわち、ファイバチャネル６８のデータ転送可能距離を越える場合には、ファイバチャネル６８に加え、エクステンダ装置を介してATMなどの広域回線を経由して、各ディスクアレイ装置２００間を接続しても良い。

【００１９】

リモートコンソール４０も、ＣＰＵや主記憶装置を有する計算機である。リモートコンソール４０、プライマリホスト１００Ａ、セカンダリホスト１００Ｂ、プライマリディスクアレイ装置２００Ａ及びセカンダリディスクアレイ装置２００Ｂは、LAN(Local Area Network)やWAN(Wide Area Network)などのIPネットワーク４８を介して相互に接続される。すなわち、データ処理システム５０はプライマリストレージシステムとセカンダリストレージシステムを接続する少なくとも２つの通信リンク、すなわちディスクアレイデバイス２００Ａと２００Ｂを接続するファイバチャネル６８と、ホスト１００Ａと１００Ｂを接続するＩＰネットワーク４８を持つ。本実施形態ではこれらの２つの通信リンクは上に述べたように異なる技術としているが、同じ技術（たとえば共にＩＰネットワーク）としても良い。

【００２０】

図１０は、図１のデータ処理システムの論理的な構成を示す図である。

【００２１】

各ホスト１００では、各ディスクアレイ装置２００間のデータ転送を制御するプログラムである非同期コピーマネージャ１５０がＣＰＵ１１０で実行される。非同期コピーマネージャ１５０は、主記憶装置１２０に格納される。非同期コピーマネージャ１５０を実行する各ホスト１００は、ホストに直接入力されるまたはあらかじめスケジューリングされたユーザの要求を受けて、各ディスクアレイ装置２００が実行するジャーナル処理（ジャーナルの取得処理、ジャーナルの転送処理、及びジャーナルを用いたデータの復元処理）の管理を行なう。詳細は後

述する。

【００２２】

また、ジャーナル処理中は各ホスト１００の非同期マネージャ１５０間でIPネットワーク４８を用いて随時通信が行われ、ジャーナル処理に必要な管理情報、具体的には、後述するジャーナル作成状況等が交換される。

【００２３】

各ディスクアレイ装置２００の記憶制御装置２１０では、コピープログラム２１１０及びジャーナルバックアップ・リストアプログラム２１２０が、プロセッサ２１４で実行される。これらのプログラムは制御メモリ２１５に格納される。ジャーナルバックアップ・リストアプログラム２１２０は、ジャーナルバックアッププログラムとジャーナルリストアプログラムから構成される。又、記憶制御装置２１０は、コピープログラム２１１０及びジャーナルバックアップ・リストアプログラム２１２０の他に、各ホストからの指示に基づいて、ディスク装置２２０への入出力処理を実行する。

【００２４】

ディスク装置２２０には、１つあるいは複数の論理記憶領域（ボリューム）が作成または関連付けられる。これら複数の論理ボリュームは、ユーザの指定により、データボリューム領域２２１０とジャーナルボリューム領域２２２２として使用される。各ホスト１００は、非同期コピーマネージャ１５０を実行することで、各ディスクアレイ装置２００が有するジャーナルバックアップ・リストアプログラム２１２０とコピープログラム２１１０の実行を制御する。尚、各ホスト１００では、ユーザが使用するアプリケーションプログラム１５２や、ディスクアレイ装置制御インターフェースとなるプログラム（以下「RAIDマネージャ」）も、そのホスト１００が有するＣＰＵ１１０で実行される。又、非同期コピーマネージャ１５０とRAIDマネージャはプログラム間通信を用いて相互に情報を遣り取りする。

【００２５】

リモートコンソール４０では、リモートコンソールストレージナビゲータと称するプログラム４２が実行される。リモートコンソール４０は、プログラム４２

を実行することによって、本実施形態のデータ処理システムの各構成要素、具体的には、各ホスト１００や各ディスクアレイ装置２００の管理を行なう。プログラム４２は、リモートコンソール４０の主記憶装置に格納される。

【００２６】

尚、ここまでに説明した各プログラムは、コンパクトディスクや光磁気ディスクといった可搬記憶媒体を用いて、あるいは、IPネットワーク４８を介して、他の装置から各装置が有する記憶媒体にインストールされる。

【００２７】

図２は、第一の実施形態のデータ処理システムの動作の概略を示すフローチャートである。

【００２８】

ステップ９１００でユーザは、ホスト１００あるいはリモートコンソール４０が持つGUI(graphical user interface)を用いてホスト１００（プライマリホスト１００Ａでもセカンダリホスト１００Ｂでも良い）にペア生成コマンドを入力する。ペア生成コマンドは、ジャーナル取得の対象となる、プライマリディスクアレイ装置２００Ａが有する第一のボリューム（以下「PVOL」）２２１２と、PVOLに格納されるデータの複製先（ペア）となる、セカンダリディスクアレイ装置２００Ｂが有する第二のボリューム（以下「SVOL」）２２１４とを関連付けるコマンドである。

【００２９】

ペア生成コマンドを受けたホスト１００は、プライマリディスクアレイ装置２００Ａにおいて、指定されたPVOL２２１２に対応するジャーナルを格納するボリューム（以下「ジャーナルボリューム」）２２２２Ａを割り当て、またセカンダリディスクアレイ装置２００Ｂにおいて、指定されたSVOL２２１４に対応するジャーナルを格納するジャーナルボリューム２２２２Ｂを割り当てるよう、各ディスクアレイ装置２００を制御する（ステップ９１１０）。ここで、PVOL２２１２とそれに割り当てられたジャーナルボリューム２２２２Ａのペアをジャーナルグループと呼び、SVOL２２１４とそれに割り当てられたジャーナルボリューム２２２２Ｂのペアもジャーナルグループと呼ぶ。ジャーナルグループは「ジャーナル

ペア」と呼ばれることもある。又、ペア生成コマンドはPVOLのジャーナルペアとSVOLのジャーナルペアも関連づける。この関連づけ（すなわちジャーナルペアのペア）を「デバイスグループ」と呼ぶ。

【００３０】

尚、PVOL及びSVOLには、記憶容量を拡張するために、単一のボリュームに限らず複数のボリュームの集合（ボリュームグループ）を割り当てることも出来る。この場合、ユーザはペア生成コマンド入力時に、ボリュームグループを指定する。各ディスクアレイ装置２００は、指定されたボリュームグループを仮想的な１つのボリュームとして扱い、単一のボリュームに対するのと同様にジャーナル処理を行なう。又、ジャーナルボリュームも同様にボリュームグループとしても良い。

【００３１】

尚、本実施形態においては、プライマリホスト１００Ａがプライマリディスクアレイ装置２００Ａを主に制御し、セカンダリホスト１００Ｂがセカンダリディスクアレイ装置２００Ｂを主に制御する。したがって、セカンダリホスト１００Ｂにペア生成コマンドが入力された場合、セカンダリホスト１００Ｂは、ペア生成コマンドに含まれる情報（デバイスグループを指定する情報）のうち、プライマリディスクアレイ装置２００Ａにて必要な情報を、ＩＰネットワーク４８を介してプライマリホスト１００Ａに転送する。同様に、プライマリホスト１００Ａにペア生成コマンドが入力された場合は、セカンダリディスクアレイ装置２００Ｂにて必要な情報を、セカンダリホスト１００Ｂに転送する。

【００３２】

ジャーナルボリュームの割り当て方法としては、少なくとも次の２つが本実施形態で可能である。

（１）ペア生成コマンド入力時に、ユーザ自身がジャーナルボリュームを指定する。

（２）ホスト１００が未使用の論理ボリュームを任意に選択して使用する。たとえば、以下の方法が取られる。まず、予め、各ディスクアレイ装置２００において、未使用の論理ボリュームが制御メモリ２１５内にてジャーナルボリュームプ

ールとして管理される。

【００３３】

各ディスクアレイ装置２００は、ジャーナルボリュームプール内に登録されている論理ボリュームに関する情報、例えばディスクアレイ装置２００Ａ内の物理アドレス、物理ボリュームの容量等を各ホスト１００に通知する。ペア生成コマンドが入力されたホスト１００は、この通知された情報に基づいて、適当なボリューム（単数でも、複数でも構わない。複数の場合、ホスト１００はその複数のボリュームを１つの仮想的ボリュームとして扱う）をジャーナルボリュームとして選択する。

【００３４】

尚、ユーザは、ホスト１００で実行されている通常のアプリケーションが、ジャーナルボリュームに対し入出力要求を発行できるかどうかを指定出来る。これには、（１）通常の入出力処理に使用されるディスク装置２２０上に割り当てられたボリュームをジャーナルボリュームとして選択する場合と、（２）ホスト１００が通常の入出力処理に使用できないボリュームをジャーナルボリュームとして選択する場合がある。

【００３５】

前者の場合、ホスト１００の通常のアプリケーションからジャーナルの参照が可能である。また、ディスクアレイ装置２００にファイバチャネルで接続された別のホスト１００のアプリケーションからジャーナルの参照も可能である。従って、通常のアプリケーションで、ジャーナルに関する統計情報の取得や管理を行なうことが可能であるが、誤ってジャーナルを破壊してしまう可能性もある。

【００３６】

一方、後者の場合、非同期コピーマネージャ１５０を実行するホスト１００が各ディスクアレイ装置２００にジャーナルコピー・リストアプログラム又はコピープログラムを実行させる場合にのみ、ホスト１００のジャーナルの参照が許されることとなる。したがって、ホスト１００からの通常の入出力処理でジャーナルが破壊されることが無い。

【００３７】

その後、ジャーナル処理が、たとえばプライマリストレージシステム１０で、実行される（ステップ９１２０）。ジャーナル処理は、取得処理９１２２、コピー処理９１２４及びリストア処理９１２６を含む。プライマリディスクアレイ装置２００Ａは、プライマリホスト１００Ａからジャーナルの取得を指示するコマンド（以下「ジャーナル取得開始コマンド」）を受信したことを契機に、ジャーナルの取得処理９１２２を開始する。ジャーナル取得開始コマンド受領後、プライマリディスクアレイ装置２００Ａは、PVOL２２１２へのデータの各書き込み処理（ステップ９２００、９２１０）の後、ジャーナルデータ及びメタデータをジャーナルボリューム２２２２Ａに格納する(ステップ９２２０)。尚、ジャーナルデータは書き込み処理による更新後データ（つまり、書き込まれたデータ）のコピーであり、メタデータは、更新データがPVOL２２１２に書込まれた時刻、更新データの格納アドレス、対応するジャーナルデータのジャーナルデータ領域への格納アドレスやデータ長に関する情報である。ジャーナルはジャーナルデータとそれに対応するメタデータから構成される。

【００３８】

ジャーナル取得処理９１２２が開始された後、各ホスト１００は、非同期コピーマネージャ１５０を実行して、定期的にジャーナルコピー処理９１２４を制御する。ジャーナルコピー処理９１２４とは、ディスクアレイ装置２００間でジャーナルの転送を行う処理である。ジャーナルコピー処理９１２４は、プライマリホスト１００Ａが、プライマリディスクアレイ装置２００Ａから得たジャーナル作成状況に関する情報（詳細は後述）に基づいてジャーナルをコピーする必要があると決定した時に開始される（たとえば、予め決められた量の情報がプライマリディスクアレイ装置２００Ａのジャーナルボリューム２２２２Ａに格納された時）。次にプライマリホスト１００Ａは、セカンダリホスト１００Ｂにリンク４８を経由して通知する。これにより、セカンダリホスト１００Ｂがセカンダリディスクアレイ装置２００Ｂにデータ読み出しの要求（以下、「ジャーナルコピー要求コマンド」）を発行して、プライマリディスクアレイ装置２００Ａからのジャーナル転送を開始させる（ステップ９３００）。

【００３９】

ジャーナルコピー要求コマンドを受取ったセカンダリディスクアレイ装置２００Ｂは、プライマリディスクアレイ装置２００Ａに対して、データの読み出し要求を発行する（ステップ９３１０）。プライマリディスクアレイ装置２００Ａは、要求されたデータを、コピープログラム２１１０を実行することで、セカンダリディスクアレイ装置２００Ｂに送信する。ジャーナルコピー処理９１２４の詳細は後述する。

【００４０】

一方、ジャーナル取得処理９１２２が開始される前にPVOL２２１２に格納されていたデータは、ジャーナルコピー処理９１２４が開始されてもセカンダリディスクアレイ装置２００Ｂには転送されない。したがって、別途PVOL２２１２からSVOL２２１４へこれらのデータ（以下「イニシャルデータ」）をコピーする必要がある。そのため、本実施形態においては、イニシャルデータをPVOL２２１２からSVOL２２１４に転送する形成コピー処理が実行される（ステップ９１３０）。イニシャルデータは、ホスト１００の指示に基づいて、PVOL２２１２のボリューム先頭領域から末尾まで順に転送される。本処理も、各ディスクアレイ装置２００のコピープログラム２１１０に基づいて実行される。

【００４１】

形成コピーとジャーナルコピー処理９１２４は、非同期かつ並行に行なわれる。つまり、形成コピーは、ペア生成コマンドに基づいてPVOL２２１２及びSVOL２２１４が指定された後であれば、ジャーナル取得処理９１２２及びジャーナルコピー処理９１２４の実行前でも実行中でも行うことが出来る。ただし、形成コピーが完了しない限り、セカンダリディスクアレイ装置２００Ｂにおいてリストア処理９１２６を行なっても、SVOL２２１４にはPVOL２２１２の内容が完全には反映されない。リストア処理９１２６は、コピー処理９１２４に従いプライマリディスクアレイ装置２００Ａから受取ったジャーナルを用いて、ＳＶＯＬ２２１４へ、ＰＶＯＬ２２１２のデータの更新を反映することを含む。

【００４２】

尚、形成コピーを、セカンダリディスクアレイ装置２００Ｂがプライマリディスクアレイ装置２００Ａにデータを要求する１つまたは複数のリードコマンドを

発行することによって実行して、プライマリディスクアレイ装置２００Ａの負担を軽減できる。

【００４３】

すべてのイニシャルデータがセカンダリディスクアレイ装置２００ＢのSVOL２２１４へコピーされたら、コピープログラム２１１０Ｂは、セカンダリホスト１００Ｂへ形成コピー終了を報告する。それ以後セカンダリサイト２０でのデータの正確なリカバリが可能になる。一般に形成コピーはジャーナル取得処理開始以後に開始される。

【００４４】

ジャーナル取得処理９１２２は、ホスト１００Ａが、プライマリディスクアレイ装置２００Ａへのジャーナル取得の停止を指示するコマンド（以下「ジャーナル取得停止コマンド」）の発行を指示することで終了する。

【００４５】

又、ホスト１００Ｂからのコマンド（以下「ジャーナルリストア要求コマンド」）に応じて（ステップ９４００）、セカンダリディスクアレイ装置２００Ｂは、ジャーナルボリューム２２２２Ｂに格納されたジャーナルを用いてＳＶＯＬ２２１４に格納されたデータをリストアする（ステップ９４１０）。この処理をジャーナルリストア処理９１２６と言う。ジャーナルリストア処理９１２６の詳細は後述する。

【００４６】

図３は、本発明の計算機システムの第一の実施形態におけるジャーナル取得９１２２、ジャーナルコピー処理９１２４、及びジャーナルリストア処理９１２６の動作を示す図である。ホスト１００Ａ及び１００Ｂでそれぞれ非同期コピーマネージャ１５０を実行して、これらの処理を制御する。プライマリディスクアレイ装置２００Ａは、ジャーナルバックアップ・リストアプログラム２１２０のうちジャーナルバックアッププログラム２１２２を実行する。ジャーナルバックアッププログラム２１２２を実行することによって、プライマリディスクアレイ装置２００Ａは、PVOL２２１２に書き込まれるデータのコピーをジャーナルデータとして、ジャーナルボリューム２２２２Ａへ格納する。又、プライマリディスク

アレイ装置２００Ａはジャーナルの一部としてメタデータもジャーナルボリューム２２２２Ａへ格納する。上記のステップが、ジャーナル取得処理９１２２である。

【００４７】

一方、セカンダリディスクアレイ装置２００Ｂは、ジャーナルバックアップ・リストアプログラム２１２０のうち、ジャーナルリストアプログラム２１２４を実行してジャーナルリストア処理９１２６を行なう。ジャーナルリストアプログラム２１２４は、ジャーナルボリューム２２２２Ｂに格納されたジャーナルをリストアして、PVOL２２１２で更新されたデータをデータボリューム２２１４へ反映する。

【００４８】

以下、図３を用いて、ジャーナル取得９１２２、ジャーナルコピー処理９１２４及びジャーナルリストア処理９１２６の手順を説明する。

【００４９】

プライマリディスクアレイ装置２００Ａでデータボリューム２２１０のジャーナル取得処理９１２２が開始されると、プライマリディスクアレイ装置２００Ａは、プライマリホスト１００AからPVOL２２１２への書き込み処理５１００に応じて、ジャーナルを作成し、作成したジャーナルをジャーナルボリューム２２２２Ａへ格納する。（ステップ５２００）。

【００５０】

プライマリホスト１００Ａは、非同期コピーマネージャ１５０を実行して特定のコマンド（以下「ジャーナル作成状況取得コマンド」）を発行することで、プライマリディスクアレイ装置２００Ａから、ジャーナル作成状況に関する情報（たとえばジャーナルボリューム内のジャーナルの容量）を取得する。（ステップ５３００）。

【００５１】

プライマリホスト１００Ａが取得したジャーナル作成状況に関する情報は、ホスト間で処理を連携するために、IPネットワーク４８を経由してセカンダリホスト１００Ｂへ通知される（ステップ５０００）。この情報は、具体的には、ジャ

ーナルボリューム２２２２Ａがセカンダリディスクアレイ装置２００Ｂへコピーできるようになった時にホスト１００Ａからホスト１００Ｂへ通知する際に使用される。

【００５２】

セカンダリホスト１００Ｂは、非同期コピーマネージャ１５０を実行し、GUIを介したユーザからの指示の入力あるいは予め定められていたスケジュール（たとえばプライマリディスクアレイ装置２００Ａで一定量のジャーナルがジャーナルボリュームに格納された時、または一定期間ごと）に従い、セカンダリディスクアレイ装置２００Ｂに対し、ジャーナルコピー要求コマンドを発行する（ステップ５４００）。

【００５３】

ジャーナルコピー要求コマンドには、コピーすべきジャーナル（複数でも良い）、そのジャーナルが格納されているジャーナルボリューム、そのジャーナルボリュームを有するディスクアレイ装置２００を指定する情報、及びコピーしたジャーナルが格納されるジャーナルボリュームを指定する情報が含まれている。

【００５４】

ジャーナルコピー要求コマンドを受信したセカンダリディスクアレイ装置２００Ｂの記憶制御装置２１０は、コピープログラムを実行することで、リードコマンドをプライマリディスクアレイ装置２００Ａに対して発行する（ステップ５５００）。このリードコマンドを受取ったプライマリディスクアレイ装置２００Ａは、リードコマンドで指定されたジャーナルをセカンダリディスクアレイ装置２００Ｂに送信する（ステップ５６００）。セカンダリディスクアレイ装置２００Ｂに送信されたジャーナルが格納されていた領域はパージ（開放）され、新たなジャーナルの格納に利用される。

【００５５】

ジャーナルを受信したセカンダリディスクアレイ装置２００Ｂは、ジャーナルコピー要求コマンドで指定されたジャーナルボリューム２２２２Ｂに、受信したジャーナルを格納する。その後、セカンダリホスト１００Ｂは、セカンダリディスクアレイ装置２００Ｂに対して、ジャーナルリストア要求コマンドを発行する

（ステップ５７００）。

【００５６】

ジャーナルリストア要求コマンドを受信したセカンダリディスクアレイ装置２００Ｂは、ジャーナルリストアプログラム２１２４を実行して、ジャーナルボリューム２２２２ＢからSVOL２２１４へデータのリストアを行なう（ステップ５８００）。尚、リストアが終わったジャーナルが格納されていた領域はパージ（開放）され、新たなジャーナルの格納に利用される。

【００５７】

尚、非同期コピーマネージャ１５０を実行するホスト１００は、ホストフェイルオーバを実行することが出来る。具体的には、プライマリホスト１００Ａが何らかの理由で利用不能になり、ジャーナルコピー処理の継続が不可能になった場合、セカンダリホスト１００Ｂがプライマリホスト１００Ａの機能を併せて実行する。

又、プライマリディスクアレイ装置が、ストレージエリアネットワークで複数のプライマリホストに接続されていても良い。こうした環境でも前記ジャーナル取得や他の処理は、当業者によって実行できるよう容易に修正可能である。

【００５８】

図４は、本実施形態で用いられるPVOL２２１２とジャーナルボリューム２２２２Ａの対応を示す図である。以下、ジャーナルボリューム２２２２Ａを正ジャーナルボリュームと呼び、ジャーナルボリューム２２２２Ｂを副ジャーナルボリュームと呼ぶ。双方のデータ構造は基本的に同一である。

【００５９】

一般に、PVOL、SVOLおよびジャーナルボリュームは各々予め定められた論理ブロック単位で管理される（たとえば５１２ＫＢ）。論理ブロックの各々には、論理ブロックアドレス（以下「LBA」）が付与されている。

【００６０】

正ジャーナルボリュームは、メタデータ領域７１００及びジャーナルデータ領域７２００を有する。ジャーナルデータ領域７２００には、先に説明したジャーナルデータ７２１０、即ち、ライトコマンドによってPVOLに書き込まれたデータ

５１１０のコピーが格納される。メタデータ領域７１００には、先に説明したメタデータ７１１０、即ち、PVOL２２１２の更新が行われた時刻、更新データの格納アドレス７１１２、対応するジャーナルデータ７２１０のジャーナルデータ領域の格納アドレス７１１４及び更新データ長が格納される。

【００６１】

各アドレスはLBAで、データ長は論理ブロック数で各々表されても良い。また、データが格納されている場所は、データが格納された領域（ジャーナルデータ領域又はメタデータ領域）のベースアドレス（先頭LBA）との差分（オフセット）で表されても良い。本実施形態において、メタデータのデータ長は一定（例えば６４バイト）であるが、ジャーナルデータのデータ長は、ライトコマンドで更新されるデータに依存するので一定ではない。

【００６２】

ジャーナルグループ定義時に、各ディスクアレイ装置２００は、設定されるジャーナルボリューム２２２２に対して、メタデータ領域７１００及びジャーナルデータ領域７２００の設定を行う。具体的には、各領域の先頭LBA及びブロック数が指定される。各ホスト１００は、非同期コピーマネージャ１５０を実行して、設定された領域に関する情報（先頭LBA、ブロック数）を要求するコマンド（ジャーナルグループ構成取得コマンド）をディスクアレイ装置２００に発行する。これにより、各ホスト１００は、各ディスクアレイ装置２００が設定したメタデータ領域７１００及びジャーナルデータ領域７２００に関する情報を取得することができる。

【００６３】

図１７は、本実施形態で用いられるジャーナルボリューム２２２２ＢとSVOL２２１４の対応を示す図である。副ジャーナルボリュームは、正ジャーナルボリュームと同様にメタデータ領域７１００及びジャーナルデータ領域７２００を有する。メタデータ領域７１００には、正ジャーナルボリュームのメタデータ領域から転送されたメタデータ７１１０Ｂが格納される。ジャーナルデータ領域７２００には、正ジャーナルボリューム２２２２Ａのジャーナルデータ領域から転送されたジャーナルデータ７２１０Ｂ（メタデータ７１１０Ｂに対応する）が格納さ

れる。

【００６４】

メタデータはPVOL２１１２で行われたデータ更新の情報を持ち、そのアドレス情報７１１４は対応するジャーナルデータ７２１０（副ボリュームのジャーナルデータ領域へコピーされる）のアドレスを示す。更に、ジャーナルデータ７２１０を、副ジャーナルボリューム２２２２Ｂのジャーナルデータ領域７２００からアドレス７１１２に対応するSVOL２２１４のアドレスへコピーすることによって、PVOL２２１２での更新をSVOL２２１４へ反映できる。

【００６５】

図５は、本実施形態の正ジャーナルボリュームと副ジャーナルボリュームのジャーナルデータ領域を示す図である。

【００６６】

正ジャーナルボリュームと副ジャーナルボリュームはそれぞれLBAでアドレスづけられていて、正副ジャーナルボリュームの各々のLBAは１対１に対応付けられている。

【００６７】

正ジャーナルボリュームが有するジャーナルデータ領域７２００は、ジャーナルデータが格納されているジャーナル格納済み領域２２３２、２２３３、及び２２３４と、ジャーナルデータが格納されていないパージ済み領域２２３１とに区別される。パージ済み領域は、PVOL２２１２の新たなジャーナルデータの格納に使用される。

【００６８】

副ジャーナルボリュームが有するジャーナルデータ領域７２００は、既にSVOLへのリストアに使用されたジャーナルデータが格納されている（又はジャーナルデータが格納されていない）リストア済み領域４２３１、SVOLへのジャーナルリストアの対象として指定されたジャーナルデータが格納されているリストア中領域４２３２、ジャーナルリストアの対象となっていないジャーナルデータが格納されているリード済み領域４２３３、及び正ジャーナルボリュームから転送中のジャーナルデータが格納されるリード中領域４２３４とに区別される。

【００６９】

正ジャーナルボリュームのパージ済み領域２２３１は、副ジャーナルボリュームのリストア中領域４２３２又はリストア済み領域４２３１の一部と対応づけられる。

【００７０】

正ジャーナルボリュームのジャーナル格納済み領域は、副ジャーナルボリュームのリード済み領域４２３３、リード中領域４２３４又はリストア済み領域４２３１の一部と対応づけられる。ここで、リード済み領域４２３３に対応するジャーナル格納済み領域２２３２は、ジャーナルがセカンダリディスクアレイ装置２００Ｂへ送信された後であるので、パージ可能である。又、リード中領域４２３４に対応するジャーナル格納済み領域２２３３に格納されたジャーナルデータは、データ転送の対象となっているために、パージすることが出来ない。ジャーナル格納済み領域２２３２は、対応するジャーナルが転送された後直ちにパージする必要は無い。パージは定期的に行なっても良いし、プライマリホスト１００Ａからの指示（「ジャーナルパージコマンド」）に従ってパージしても良い。

【００７１】

正副ジャーナルボリュームのジャーナルデータ領域７２００が有する各領域は、各領域の境界にある論理ブロックのLBAを示すポインタで各ホスト１００に識別される。プライマリホスト１００Ａがプライマリディスクアレイ装置２００Ａから取得するジャーナル処理状況に関する情報とは、具体的にはこれらポインタの値である。

【００７２】

各ホスト１００は、接続されたディスクアレイ装置２００にジャーナル作成状況取得コマンドを発行することによって、これらポインタの値をディスクアレイ装置２００から取得する。その後、ホスト１００は、これらポインタの値を用いて、ジャーナルボリュームのどの領域にジャーナルデータが格納されているかを判断する。尚、これらポインタの値は、制御メモリ２１５に格納されていても良い。

【００７３】

以下、各ポインタについて説明する。尚、図５において、LBAは、図の上から下に向けて番号が割り振られている。したがって、図の一番上のLBAの番号が一番小さい。また、ジャーナルボリュームはサイクリックバッファと同じように、繰り返し使用される。つまり、ジャーナルボリュームの末尾の論理ブロックまで使用したら、先頭の論理ブロックが再度使用される。正副ジャーナルボリュームのいずれでも、ジャーナルは作成された順番にジャーナルボリュームに書き込まれる。まず、正ジャーナルボリュームのポインタについて説明する。

【００７４】

ジャーナルアウトLBA２２４１は、ジャーナル格納済み領域先頭の論理ブロックに対応するLBAを示すポインタである。このポインタで示される論理ブロックに、パージされていないもっとも古いジャーナルデータが格納されている。プライマリホスト１００Ａ又はセカンダリホスト１００Ｂは、ジャーナルアウトLBA２２４１で示されるLBAに対応する論理ブロックを、転送対象となるジャーナルデータの先頭の論理ブロックとして認識する。

【００７５】

ジャーナルインLBA２２４２は、ジャーナルデータが格納されている末尾の論理ブロックに隣接する空の論理ブロックに対応するLBA、即ち、次にジャーナル取得処理によってジャーナルデータが格納され始める論理ブロックに対応するLBAを示すポインタである。プライマリホスト１００Ａ又はセカンダリホスト１００Ｂは、ジャーナルインLBA２４２２で示されるLBA以上のLBAを有する論理ブロックがジャーナルデータ格納のために使用可能であると認識する。

【００７６】

又、プライマリホスト１００Ａ又はセカンダリホスト１００Ｂは、ジャーナルアウトLBA２２４１で示されるLBAから、ジャーナルインLBA２２４２で示されるLBAの直前までの領域に、ジャーナルデータが格納されていると認識する。したがって、ジャーナルアウトLBA=ジャーナルインLBAならば、正ジャーナルボリュームのジャーナルデータ領域には、副ジャーナルボリュームに転送すべきジャーナルデータは含まれていないとプライマリホスト１００Ａ又はセカンダリホスト１００Ｂは判断する。

【００７７】

次に、副ジャーナルボリュームのポインタについて説明する。

【００７８】

リストア済みLBA４２４１は、リストア処理が完了した論理ブロックのうち、最も大きなLBAを有する論理ブロックを示すポインタである。したがって、リストア済みLBAポインタで示されるLBAより小さなLBAを有する論理ブロックは、新たに正ジャーナルボリュームから転送されたジャーナルデータの格納に使用される。即ち、リストア済みLBA４２４１以下のLBAを有する論理ブロックでは、ジャーナルデータがパージされている。

【００７９】

尚、副ジャーナルボリュームのパージ処理は、リストア処理終了後、記憶制御装置２１０が自動的に行なって良い。又、ジャーナルデータのパージは、実際にジャーナルデータを無意味なデータで上書きすること、又は単にポインタを移動してその領域をライト（上書き）可能とすることでも実現できる。PVOLのジャーナル格納済み領域２２３２のパージと同様に、副ジャーナルボリュームのパージもリストア処理終了後即座に行なう必要はない。

【００８０】

リストア予定LBA４２４２は、リストア済みLBA４２４１より一つ大きいLBAで示される論理ブロックから、リストア予定LBA４２４２で示される論理ブロックまでに格納されているジャーナルデータを用いて、SVOLをリストアするジャーナルリストアコマンドがセカンダリホスト１００Ｂから出されていることを示すポインタである。したがって、リストア予定LBA=リストア済みLBAならば、副ジャーナルボリュームには、リストア対象のジャーナルデータが存在しない。

【００８１】

リード済みLBA４２４３は、プライマリディスクアレイ装置２００Ａから受取ったジャーナルデータが格納されている論理ブロックのうち、最も大きいLBAを有する論理ブロックを指すポインタである。言い換えると、本ポインタは、セカンダリディスクアレイ装置２００Ｂがプライマリディスクアレイ装置２００Ａに対して最後に発行したリードコマンドに基づいて転送されたジャーナルデータの

末尾が格納されている論理ブロックを示す。

【００８２】

セカンダリホスト１００Ｂは、リード済みLBA４２４３ポインタによって、本ポインタで示されるLBAに格納されたジャーナルデータに対応する正ジャーナルボリュームのジャーナルデータまでが、副ジャーナルボリュームに格納されたことを確認する。その確認を行ったセカンダリホスト１００Ｂは、プライマリホスト１００Ａに対して、リード済みLBA４２４３についての情報を通知する。プライマリホスト１００Ａは、通知された情報に基づいて、リード済みLBA４２４３に対応するジャーナルデータが格納された論理ブロックまでジャーナルデータ領域をパージするよう、プライマリディスクアレイ装置２００Ａに指示できる。尚、ここでのパージも、ジャーナルアウトLBA２２４１ポインタの移動で実現されても良い。

【００８３】

リード予定LBA４２４４は、セカンダリホスト１００Ｂがセカンダリディスクアレイ装置２００Ｂに発行した最新のジャーナルコピー要求の対象となったジャーナルデータ領域の末尾の論理ブロックのLBAを示すポインタである。したがって、リード予定LBA=リード済みLBAならば、ジャーナルコピーの対象となるジャーナルデータが存在しない。つまり、各ディスクアレイ装置２００が実行中のジャーナルコピー処理は無い。

尚、正副ジャーナルボリュームのメタデータ領域同士も図５と同様の対応関係がある。ジャーナルデータ領域と同様に、メタデータ領域用の各ポインタ（ジャーナルアウトLBA、ジャーナルインLBA、リストア済みLBA、リストア予定LBA,リード済みLBA、リード予定LBA。これらはジャーナルデータ領域用のポインタとは別個のポインタである）で各ホスト１００および記憶制御装置２１０から制御される。

【００８４】

各ホスト１００は、非同期コピーマネージャ１５０を双方で実行することで、各ディスクアレイ装置２００におけるジャーナル取得状況を、各ポインタの値を取得することで確認する。たとえば、各ホスト１００は、正副ジャーナルボリュ

ームについて、ジャーナルペア指定時に決定されたジャーナルボリュームの大きさと、ディスクアレイ装置２００から取得したポインタの値の差分とから、ジャーナルがジャーナルボリュームにどの程度,例えば何パーセント蓄積されているか算出する。

【００８５】

そして、各ホスト１００は、算出結果に基づいて、正ジャーナルボリュームに格納されたジャーナルをどこまでパージするか、正ジャーナルボリュームに格納されたジャーナルのうち、何処までをセカンダリディスクアレイ装置２００Ｂへ転送するか、転送されたジャーナルのうち、どこまでをSVOLにリストアするか、等を各ディスクアレイ装置２００に対して指示する。

【００８６】

例えば、セカンダリホスト１００Ａは、正ジャーナルボリュームの容量の５０パーセント以上にジャーナルが格納されたことを判断した場合に、セカンダリディスクアレイ装置２００Ｂに対してジャーナルコピー要求を発行するとしても良い。

【００８７】

ここで、各ホスト１００が各ディスクアレイ装置２００に対して行う指示とは、具体的には、ジャーナル作成状況取得コマンドと、ジャーナル処理コマンドである。

【００８８】

ジャーナル作成状況取得コマンドは、正ジャーナルボリュームにジャーナルがどの程度蓄積されたかの情報を取得するためにプライマリホスト１００Ａから発行される場合と、副ジャーナルボリュームのリード処理およびリストア処理がどこまで進んでいるかの情報を取得するためにセカンダリホスト１００Ｂから発行される場合がある。

【００８９】

ジャーナル処理コマンドは、ジャーナルのパージをプライマリディスクアレイ装置２００Ａに実行させるためにプライマリホスト１００Ａから発行される場合と、ジャーナルコピー処理およびジャーナルリストア処理をセカンダリディスク

アレイ装置２００Ｂに実行させるためにセカンダリホスト１００Ｂから発行される場合とがある。従って、ジャーナルコピー要求コマンド及びジャーナルリストア要求コマンドは、ジャーナル処理コマンドの一種である。

【００９０】

尚、正ジャーナルボリュームと副ジャーナルボリュームのLBAは１対１対応としたが、正ジャーナルボリュームのLBAから副ジャーナルボリュームのLBAへの適当なアドレス変換手段を用いれば、副ジャーナルボリュームの領域が正ジャーナルボリュームより広く指定されても良い。

【００９１】

図６は、ジャーナル取得処理９１２２、ジャーナルコピー処理９１２４並びにリストア処理９１２６の詳細を示すフローチャートである。

【００９２】

プライマリホスト１００Ａは、あらかじめユーザの要求によってスケジューリングされた間隔、又は所定の間隔で、定期的にプライマリディスクアレイ装置２００Ａのジャーナル格納済み領域に関する情報をジャーナルアウトLBA及びジャーナルインLBAのポインタを用いて取得し（ステップ６１００、６２００。図３のステップ５３００）、取得した情報をセカンダリホスト１００Ｂへ送る（ステップ６１１０）。

【００９３】

セカンダリホストは、通知されたジャーナル格納済み領域を示す情報に基づいて、ジャーナルコピー処理の対象となる正ジャーナルボリュームの論理ブロック領域を決定する。尚、ジャーナルコピーの対象となる正ジャーナルボリュームの論理ブロック領域は、プライマリホスト１００Ａが予め決定することもできる。

【００９４】

その後、セカンダリホスト１００Ｂは、セカンダリディスクアレイ装置２００Ｂに、決定された論理ブロック領域を示す情報及びジャーナルコピーの対象となるディスクアレイ装置２００を指定する情報を含んだジャーナルコピー要求コマンドを発行する（ステップ６３００。図３の５４００）。ジャーナルコピー要求を受取ったセカンダリディスクアレイ装置２００Ｂは、指定されたプライマリデ

ィスクアレイ装置２００Ａに対し、指定された論理ブロック領域に格納されたジャーナルを要求するリードコマンドを発行する。

【００９５】

なお、ジャーナル領域は、図４で示すようにメタデータ領域とジャーナルデータ領域に分かれるので、ジャーナルコピー要求ではまずメタデータ領域が指定される。ジャーナルコピー要求を受け取ったセカンダリディスクアレイ装置２００Ｂは、メタデータ領域の指定された論理ブロック領域をコピーするリードコマンドを発行し、読み出したメタデータから対応するジャーナルデータが格納されたジャーナルデータ領域の論理ブロック領域を決定する。その後、セカンダリディスクアレイ装置２００Ｂは、その論理ブロック領域（すなわち対応するジャーナルデータ）をコピーするリードコマンドを発行する。

【００９６】

又は、セカンダリディスクアレイ装置２００がメタデータとジャーナルデータを同時に読み出すリードコマンドを発行しても良い。その場合、各リードコマンドのアドレスとデータ長はポインタから計算される。たとえば、ジャーナルデータについては、正ジャーナルボリュームのジャーナルデータ領域内のリード予定LBA+1に対応するＬＢＡからジャーナルインLBA-1までの領域が、副ジャーナルボリュームのジャーナルデータ領域内の対応する領域にコピーされる（ステップ６４００。図３の５５００）。

【００９７】

一方、セカンダリホスト１００Ｂは、定期的にセカンダリディスクアレイ装置２００Ｂのジャーナル処理状況を、ジャーナル作成状況取得コマンドを用いて取得する（ステップ６３１０、６３２０、６４１０）。具体的には、リード済みLBA４２４３及びリード予定LBA４２４４ポインタの値をセカンダリディスクアレイ装置２００Ｂから取得する。セカンダリホスト１００Ｂは、リード済みLBA４２４３及びリード予定LBA４２４４の値が一致した場合、ジャーナルコピー処理（すなわちリード）が完了していると判断する。

【００９８】

尚、リード予定LBA４２４４の情報がセカンダリホスト１００Ｂに保持される

場合、セカンダリホスト１００Ｂは、リード済みLBA４２４３を定期的にセカンダリディスクアレイ装置２００Ｂから取得すれば、ジャーナルリード処理の完了を判断することが出来る。

【００９９】

ジャーナルコピー処理の完了を確認したセカンダリホスト１００Ｂは、ジャーナルのリードが完了した副ジャーナルボリュームの論理ブロック領域に対してリストア処理を指示するジャーナルリストア要求コマンドをセカンダリディスクアレイ装置２００Ｂに発行する。尚、副ジャーナルボリュームの容量が大きい場合や、特にユーザがデータのリストアを要求しない場合、その他、早急なリストアが要求されない場合には、リストア処理はジャーナルの転送完了後すぐに行われる必要は無く、別途、ユーザの指示等に基づいて行われる（ステップ６３３０。図３のステップ５７００）。

【０１００】

ジャーナルリストア要求コマンドを受取ったセカンダリディスクアレイ装置２００Ｂは、指定されたLBAに対応する論理ブロックに格納されたジャーナルデータを用いて、SVOLに対するリストアを実行する（ステップ６４２０、図３のステップ５８００）。

【０１０１】

又、ジャーナルコピー処理の終了を確認したセカンダリホスト１００Ｂは、リード済みLBA４２４３で示されるLBAをプライマリホスト１００Ａに通知する（ステップ６３４０）。尚、ジャーナルコピー処理の終了を検出したセカンダリホスト１００Ｂは、セカンダリディスクアレイ装置２００Ｂに対して次のジャーナルコピー要求を指示することが可能となる（ステップ６３５０）。

【０１０２】

リード済みLBA４２４３で示されるLBAを通知されたプライマリホスト１００Ａは、通知されたLBAに対応する領域までジャーナルをパージするよう、プライマリディスクアレイ装置２００Ａに指示する（ステップ６１２０）。指示を受けたプライマリディスクアレイ装置２００Ａは、指示されたLBAまで、ジャーナルボリュームをパージする（ステップ６２１０）。

【０１０３】

図７は、本発明を適用したデータ処理システム５０の第二の実施形態を示す図である。説明の便宜のため、第一の実施形態と同じ構成要素については、同じ参照番号を用いている。

【０１０４】

図７のデータ処理システム５０は、ジャーナルコピー処理において、プライマリディスクアレイ装置２００Ａがセカンダリディスクアレイ装置２００Ｂからのリードコマンドを待つのではなく、プライマリディスクアレイ装置２００Ａからセカンダリディスクアレイ装置２００Ｂに対してデータを書き込むライトコマンドを発行する点で第一の実施形態と相違する。また、セカンダリディスクアレイ装置２００Ｂの記憶制御装置２１０がジャーナルリストア処理を実行するのではなく、セカンダリホスト１００Ｂが副ジャーナルボリュームからリストアに使用するジャーナルを読み出して、SVOLのデータをリストアする（図７のステップ５９００）点も相違する。本実施形態において、ジャーナルリストアプログラムはセカンダリホスト１００Ｂ上で実行されるプログラムである。

【０１０５】

尚、プライマリストレージサブシステム１０の構成要素は、セカンダリストレージシステム２０のそれらと区別するためにプライマリデバイスまたはプライマリ構成要素と記述されるか、参照番号の後に「Ａ」付きで記述されるか、その両方である（たとえば、プライマリホスト１００、ホスト１００Ａ、プライマリホスト１００Ａ）。同様に、セカンダリデバイスの構成要素はセカンダリデバイスまたはセカンダリ構成要素と記述されるか、参照番号の後に「Ｂ」付きで記述されるか、その両方である（たとえば、セカンダリホスト１００、ホスト１００Ｂ、セカンダリホスト１００Ｂ）。

【０１０６】

本実施形態においては、ジャーナルコピー処理の主体がプライマリディスクアレイ装置２００Ａであること、ジャーナルリストア処理を行うのがセカンダリホスト１００Ｂであることから、セカンダリディスクアレイ装置２００Ｂには、特殊な機能を有しない一般的な記憶装置を用いることができる。またデータ処理シ

ステム５０はヘテロジニアスなストレージサブシステム、すなわち異なるベンダで製造された、または異なるストレージプロトコルや方法を用いたストレージサブシステムを採用可能である。

【０１０７】

プライマリサイト１０において、PVOLデータの更新（ステップ５１００）に対するジャーナル取得処理（ステップ５２００）の処理は第一の実施形態と同様である。プライマリホスト１００Ａは、プライマリディスクアレイ装置２００Ａから、ジャーナル作成状況に関する情報を随時取得する（ステップ５３００）。

【０１０８】

またプライマリホスト１００Ａは、プライマリディスクアレイ装置２００Ａに、セカンダリディスクアレイ装置２００Ｂへのジャーナルコピー要求コマンドを発行する（ステップ５４５０）。

【０１０９】

尚、ジャーナルコピー要求コマンドには、ディスクアレイ装置２００Ｂへ送信すべきジャーナルが格納されているジャーナルボリューム、ディスクアレイ装置２００Ｂを指定する情報、及びそのジャーナルをディスクアレイ装置２００Ｂで格納するべきジャーナルボリュームを指定する情報等が含まれている。

【０１１０】

ジャーナルコピー要求コマンドを受取ったプライマリディスクアレイ装置２００Ａは、ライトコマンドをセカンダリディスクアレイ装置２００Ｂに発行することで、指定されたジャーナルをセカンダリディスクアレイ装置２００Ｂに送信する（ステップ５６００）。セカンダリディスクアレイ装置２００Ｂは、プライマリディスクアレイ装置２００Ａからのライトコマンドとして受信したジャーナルを、そのコマンドで指定された副ジャーナルボリュームの領域に格納する。

【０１１１】

その後、非同期コピーマネージャ１５０を実行するセカンダリホスト１００Ｂは、副ジャーナルボリュームからジャーナルを読み出してSVOL２２１４へデータのリストアを行なう（ステップ５９００）。

【０１１２】

副ジャーナルボリュームのポインタ管理はセカンダリサイト２０のセカンダリホスト１００Ｂが行ない、ジャーナルコピー要求５４５０の作成に必要な情報(例えばコピーアドレス計算に必要なリストア済みLBA４２４１など)をプライマリホスト１００Ａに通知する。尚、リストアが終わったジャーナルが格納されていた領域はパージされ、新たなジャーナルの格納に利用される。

【０１１３】

又、本実施形態では、プライマリホスト１００Ａの指示により、プライマリディスクアレイ装置２００Ａが、PVOL２２１２のイニシャルデータを順にセカンダリディスクアレイ装置２００BのSVOL２２１４へ書き込む（ライトコマンドをセカンダリディスクアレイ装置２００Ｂに発行する）ことで形成コピーが実現する。

【０１１４】

プライマリホスト１００Ａの指示に基づいてPVOLのすべてのイニシャルデータをセカンダリディスクアレイ装置２００Ｂへ書き込んだら、コピープログラム２１１０を実行するプライマリディスクアレイ装置２００Ａは、プライマリホスト１００Ａへ形成コピーの終了を報告する。プライマリホスト１００Ａは、この報告を受信する。これ以降、セカンダリサイト２０でリストアされたSVOL２２１４は、PVOL２２１２の内容を反映するボリュームとして扱える。

【０１１５】

図８は、本発明を適用したデータ処理システムの第三の実施形態を示す図である。データ処理システム５０は、ホスト間の第一の通信リンク４８を持つが、ディスクアレイ装置２００Ａと２００Ｂ間の第二の通信リンクを持たない。外部記憶装置が第二の通信リンクの代わりに用いられる。

【０１１６】

本実施形態では、テープ装置またはそれに準ずる第一の外部記憶装置６０がプライマリホスト１００Ａに、第二の外部記憶装置６２がセカンダリホスト１００Ｂに、それぞれファイバチャネルを介して接続される。外部記憶装置６０と６２の間はファイバチャネルなどにより接続しても良いし、それら外部記憶装置が磁気テープ等の可般記憶媒体である場合には、必要に応じて装置間で記憶媒体を運

んでデータの移動を行なっても良い。

【０１１７】

本実施形態においては、プライマリディスクアレイ装置２００Ａは、第第二の実施形態と同様に、PVOL２１１２についてジャーナル取得処理を行なう。ジャーナルのコピーは、以下の（１）～（３）のステップで行われる。

（１）プライマリホスト１００Ａが第一の外部記憶装置６０に対してデータの書き込みを行なう。具体的には、プライマリホスト１００Ａは、ジャーナル取得処理開始後、GUIなどの入力装置を介したユーザの指示の入力、または予め定められたスケジュールに応じて正ジャーナルボリュームからジャーナルを読み出し、それを外部記憶装置６０に格納する。（ステップ５６２０）。

（２）第一の外部記憶装置６０に書き込まれたデータを、第二の外部記憶装置６２へ移す。本処理は、プライマリホスト１００Ａとセカンダリホスト１００Ｂのどちらが行なっても良い。データ転送の指示には、例えばＡＮＳＩ（American National Standards Institute）ＳＣＳＩ－３のＥｘｔｅｎｄｅｄ Ｃｏｐｙコマンドが用いられる。

【０１１８】

データ転送に必要なアドレス情報や、データ転送終了報告などはホスト１００Ａと１００Ｂの間の通信リンクで交換される。又、可搬記憶媒体を第一の外部記憶装置から第二の外部記憶装置に物理的に移送した後、ユーザや管理者が各ホスト１００にデータを移したことを報告しても良い（ステップ５６２２）。

【０１１９】

（３）外部記憶装置６２に格納されたデータは、セカンダリホスト１００Ｂの指示に従って、セカンダリディスクアレイ装置２００Ｂに転送される。具体的には、セカンダリホスト１００Ｂは、外部記憶装置６２に対してリードコマンド５５０５を発行し、外部記憶装置６２からジャーナルを読み出す。その後、セカンダリホスト１００Ｂは、読み出したジャーナルをもとに、第二の実施形態と同様にSVOL２２１４のデータをリストアする（ステップ５６２５）。

【０１２０】

以上の手順により、PVOLからSVOLへジャーナルボリュームを介したデータ複製

を行なうことができる。尚、形成コピーも同様に行われる。実施形態によっては、外部記憶装置６２に格納されたジャーナルはリストア処理が終了しても、特に指示が無い限り消去されない。また外部記憶装置６２には、形成コピーの結果、すなわちイニシャルデータも格納されている。

【０１２１】

さらに、ジャーナルのメタデータ内にPVOL更新時刻のタイムスタンプが含まれているので、本発明のデータ処理システムにおいては、ジャーナル取得処理開始以後任意の時刻におけるPVOLの内容を、セカンダリディスクアレイ装置２００ＢのSVOL２２１４にリストアすることできる。すなわち、PVOLの形成コピーが完了しているSVOLに対し、セカンダリホスト１００Ｂから指定された時刻以前のタイムスタンプを持つジャーナルを時間順にSVOLへ全てリストアすれば、指定時刻のPVOLの内容をSVOLに再現できる。これを、ポイント・イン・タイム・リカバリと呼ぶ。

【０１２２】

又、セカンダリディスクアレイ装置２００Ｂのユーザが指定した任意のボリューム２２１６に対して、ポイント・イン・タイム・リカバリを行うことも可能である。すなわち、外部記憶装置６２に格納されているPVOLの形成コピーの結果をまずボリューム２２１６にコピーし、その後、ボリューム２２１６へセカンダリホスト１００Ｂから指定された時刻以前の更新時刻のタイムスタンプを持つジャーナルデータを時間順にSVOLへ全てリストアする。

【０１２３】

なお、ジャーナルデータのリストアでは、同一の領域に対するジャーナルデータが複数存在する場合は、最新のタイムスタンプを持つジャーナルデータのみを使ってリストアする形式としても良い。

又、本実施形態において外部記憶装置６０と６２は同一の種類の記憶装置でも良いし別々の種類の記憶装置でも良い。またそれらは別の装置として説明したが、それらを同一の装置としても良い。

尚、第一及び第二の実施形態でも、副ジャーナルボリュームに格納されたジャーナルのうち指定した時刻以前のタイムスタンプを持つ全てのジャーナルについ

てSVOLへリストアを行なうことでポイント・イン・タイム・リカバリが実現される。しかしリストアできるPVOLの内容は、副ジャーナルボリュームに格納されている最も古いジャーナルが示すPVOL変更時刻以後のPVOLの内容に限られる。

【０１２４】

図９は、本発明を適用したデータ処理システムの第四の実施形態を示す図である。本実施形態は第三の実施形態に類似するが、外部記憶装置６０及び６２がプライマリディスクアレイ装置２００Ａ及びセカンダリディスクアレイ装置２００Ｂにそれぞれファイバチャネルで接続される点が異なる。そのため、外部記憶装置５０への形成コピーやジャーナルの格納は、プライマリホスト１００Ａからの指示５４５０に従って、プライマリディスクアレイ装置２００Ａが行なう（ステップ５６３０）。

【０１２５】

外部記憶装置６０に格納されたデータは、プライマリディスクアレイ装置２００Ａの指示５６３１に基づくデータ転送または記憶媒体の移動により外部記憶装置６２に移される（ステップ５６３２）。

【０１２６】

外部記憶装置６２からの形成コピーやジャーナルの読み出しは、セカンダリホスト１００Ｂからの指示５４００に基づき、セカンダリディスクアレイ装置２００Ｂがリードコマンド５５０７を発行することで行なわれる（ステップ５６３５）。ジャーナル取得やリストアの処理は、第一の実施形態に従う。

【０１２７】

本実施形態によっても、PVOLのデータを、ジャーナルを転送することでSVOLに非同期に複製することができる。本実施形態では、第三の実施形態と異なり、実際のデータ転送をディスクアレイ装置２００Ａと２００Ｂが行なうので、ホスト１００Ａと１００Ｂの負荷は低くなる。また、本実施形態でも、ポイント・イン・タイム・リカバリを実現することが出来る。

【０１２８】

図１１は、本発明を適用した計算機システムの第五の実施形態を示す図である。本実施形態は、先述したこれまでの実施形態と異なり、プライマリストレージ

システム１０は、複数のセカンダリストレージシステム２０及び３０と接続される。

【０１２９】

本実施形態では、プライマリディスクアレイ装置２００AのPVOL２２１２に対応するジャーナルを、セカンダリディスクアレイ装置２００BのSVOL２２１４Bに対応する副ジャーナルボリューム並びにセカンダリディスクアレイ装置２００CのSVOL２２１４Cに対応する副ジャーナルボリュームに各々転送しリストアする。又、PVOL２２１２から、SVOL２２１４B並びにSVOL２２１４Cに対し各々形成コピーも行なう。具体的には、個々のセカンダリディスクアレイ装置２００B及び２００Cが、プライマリディスクアレイ装置２００Aにリードコマンドを発行する。又は、プライマリディスクアレイ装置２００Aが、各セカンダリディスクアレイ装置２００B、Cにライトコマンドを発行する。これにより、プライマリサイトに格納されたデータの複製が複数のサイトに作成できる。

【０１３０】

図１２は、本発明を適用した計算機システムの第六の実施形態を示す図である。本実施形態においては、ユーザあるいはシステム管理者は、障害などによりセカンダリサイト２０が利用できなくなった場合に備え、予め、プライマリホスト１００Aに、セカンダリサイト２０の代わりに利用できるサイトの候補を一つあるいは複数登録しておく。こうした候補のリストまたはテーブル１６０がホスト１００A内に格納されても良い。

【０１３１】

セカンダリサイト２０が利用できなくなった場合、プライマリホスト１００Aは、リスト１６０から、新たなセカンダリサイト４０を選択する。新たなセカンダリサイトの選択の際には、プライマリホスト１００Aは、予め定められた優先順位に沿ってセカンダリサイトの選択を行っても良い。この場合、優先順位は、サイトの候補を登録する際に、ユーザが設定する場合が考えられる。あるいは、プライマリサイト１００Aと登録されたサイトとの距離、データ転送速度等の条件に基づいてプライマリサイト１００Aが自動的に登録されたサイトの選択の優先順位を設定することも考えられる。

【０１３２】

その後、プライマリホスト１００Ａは、新たに選択したセカンダリサイト４０におけるセカンダリホスト１００Ｄに対し、デバイスグループ等の情報を転送する。デバイスグループ等の情報を受取った新たなセカンダリホスト１００Ｄは、セカンダリホスト１００Ｄに接続されたディスクアレイ装置２００Ｄに対して、新たなSVOL、ジャーナルペアの設定、及び１００Ｄとプライマリホスト１００Ａとの間のジャーナルコピーを要求する。尚、ほとんどの場合、形成コピーも必要となるので、セカンダリホスト１００Ｄは、形成コピーもディスクアレイ装置２００Ｄに要求する。これらの処理により、リモートレプリケーションの前または最中にセカンダリサイト２０で障害が発生した場合にも、新たに選択されたセカンダリサイト４０において、プライマリサイト１０に格納されたデータの複製作業が続行できる。

【０１３３】

尚、利用不能になったのがセカンダリサイト２０のセカンダリホスト１００Ｂのみで、セカンダリディスクアレイ装置２００Ｂが利用可能である場合は、ディスクアレイ装置２００Ｂをセカンダリサイト４０のディスクアレイ装置として引き続き（たとえばストレージエリアネットワークシステム内で）利用することも可能である。

【０１３４】

図１３は、本発明を適用した計算機システムの第七の実施形態を示す図である。

【０１３５】

本実施形態は、先述したこれまでの実施形態と異なり、プライマリサイト１０がプライマリホスト１００Ａ及び仮想ディスクアレイ装置１５Ａから構成され、セカンダリサイト２０がセカンダリホスト１００Ｂ及び仮想ディスクアレイ装置１５Ｂから構成される。各ホスト１００は、各仮想ディスクアレイ装置１５を単一のディスクアレイ装置２００として扱う。すなわち、各ホストは、第一の実施形態と同様のコマンドを仮想ディスクアレイ装置１５に発行する。

【０１３６】

仮想ディスクアレイ装置１５は、バーチャリゼーションサーバ３００及び複数のストレージサブシステムから構成される。尚、ここでは、ストレージサブシステムの例として、ディスクアレイ装置２００を考える。また、バーチャリゼーションサーバ３００は、CPU等を有する一般の計算機である。バーチャリゼーションサーバ３００は、プライマリホスト１００Ａ（セカンダリホスト１００Ｂ）、複数の各ディスクアレイ装置２００及び他のバーチャリゼーションサーバ３００とファイバチャネルで接続される。このファイバチャネルは、第一の実施形態のファイバチャネル６６及び６８に相当し、記憶制御装置２１０間の通信や、形成コピー及びジャーナルコピー処理に用いられる。また、第一の実施形態と同様にバーチャリゼーションサーバ３００間の距離が長い場合にエクステンダ装置を介してATMなどの広域回線を経由してバーチャリゼーションサーバ３００間を接続しても良い。

【０１３７】

バーチャリゼーションサーバ３００は、バーチャリゼーションサーバ３００に接続される複数のディスクアレイ装置２００が有するボリューム（論理ボリュームでも物理ボリュームでも良い）の集合を、各ホスト１００に対し仮想的な一つ（又は複数）の記憶装置システムとして提供する。これはバーチャリゼーションサーバ３００が、バーチャリゼーションマネージャと称するプログラム３１０、すなわち接続されたホスト１００毎に、接続された各ディスクアレイ装置２００上の複数のボリュームを１つのアドレス空間を有する記憶領域（以下「仮想ストレージイメージ」）として変換する変換プログラムを実行することで実現する。

【０１３８】

ここで、ホスト１００と仮想ディスクアレイ装置１５との間のデータ転送について簡単に説明する。たとえば、ホスト１００Ａから仮想ディスクアレイ装置１５Ａの仮想ストレージイメージに書き込み要求５１００があった場合を考える。この書き込み要求５１００は、バーチャリゼーションサーバ３００Ａによって、ホスト１００Ａに対応する仮想ストレージイメージを構成する各ディスクアレイ装置２００への書き込み要求５１０５に変換される。その後、バーチャリゼーションサーバ３００Ａは、変換した書き込み要求５１０５を、各ディスクアレイ装

置２００へ送る。このとき、書き込み要求５１００に含まれる書き込みデータは、上記変換に従って、各ディスクアレイ装置２００ごとのデータに分割される。また、書き込みアドレスも各ディスクアレイ装置２００への書き込みアドレスに変換される。

【０１３９】

次に、ホスト１００Ｂから仮想ディスクアレイ装置１５Ｂの仮想ストレージイメージにデータの読み出し要求があった場合を考える。データ読み出し要求は、バーチャリゼーションサーバ３００Ｂによって、ホスト１００Ｂに対応する仮想ストレージイメージを構成する各ディスクアレイ装置２００への読み出し要求に変換される。その後、バーチャリゼーションサーバ３００Ｂは、変換した読み出し要求を、各ディスクアレイ装置２００へ送る。

【０１４０】

その後、各ディスクアレイ装置２００からバーチャリゼーションサーバ３００Ｂへ、要求されたデータが転送される（ステップ５１１５）。データを受取ったバーチャリゼーションサーバ３００Ｂは、受信したデータを統合し、ホスト１００Ｂに転送する（ステップ５１１０）。

【０１４１】

又、図１３には示していないが、各バーチャリゼーションサーバ３００は、各ホスト１００や各ディスクアレイ装置２００と同様にＩＰネットワークにてリモートコンソールに接続される。ユーザは、リモートコンソールを介してこの計算機システムを管理する。

【０１４２】

又、例えばバーチャリゼーションサーバ３００が、バーチャリゼーションサーバ３００に接続された各ディスクアレイ装置２００の入出力処理を監視することで、以下の処理を自動的に実現することが考えられる。

（Ａ）修復できるリードエラー（すなわち、リードデータにエラーが検出されるが、データと共に格納されたエラー訂正符号によって修復される）が頻発するようになったディスクアレイ装置２００を、別のディスクアレイ装置２００に置き換えるよう、マッピングを変更する。

（Ｂ）アクセス頻度の高いデータを、より高速なディスクアレイ装置２００内に配置しなおす。

【０１４３】

これらの処理に先立って本発明の技術を用いれば、バーチャリゼーションサーバ３００の制御に基づくジャーナル取得、ジャーナルコピー、及びジャーナルリストア処理により、予め置き換える元のディスクアレイ装置２００上のデータを、置き換える先のディスクアレイ装置２００上に複製することができる。その後、仮想ストレージイメージの構成を変更すれば、ホストで実行されるアプリケーションの中断無しでストレージサブシステムの追加や削除が可能である。

【０１４４】

仮想ディスクアレイ装置１５のバーチャリゼーションサーバ３００は、ジャーナルバックアップ・リストアプログラム及びコピープログラムを実行する。

【０１４５】

又、仮想ディスクアレイ装置１５は、先の実施形態で説明したようなPVOL、正副ジャーナルボリューム、又はSVOLを有する。ただし、PVOL、正副ジャーナルボリューム、並びにSVOLは、それぞれ複数のディスクアレイ装置２００にまたがるように構成されうるが、バーチャリゼーションサーバ３００によって、ホスト１００、又はバーチャリゼーションサーバ３００で実行されるジャーナルバックアップリストアプログラム及びコピープログラムからは、第一の実施形態のように（仮想的な）１つのボリュームとして扱われる。従って、バーチャリゼーションサーバ３００は、各ホスト１００の指示に従って、第一の実施形態と同様の制御、即ち、ジャーナル取得、ジャーナルコピー、ジャーナルリストア、及びジャーナルボリュームの管理を行う。

【０１４６】

尚、ユーザからの要求や予め決めらた方法に応じて、仮想化されるディスクアレイ装置２００の数を動的に増減しても良い。更に、プライマリサイト１０とセカンダリサイト２０で接続されるディスクアレイ装置２００が同数・同一種である必要は無い。本実施形態においてバーチャリゼーションサーバ３００と各ホスト１００は別の装置として説明したが、バーチャリゼーションマネージャをホス

ト１００上で動作するプログラムとすると、ホスト１００とバーチャリゼーションサーバ３００を同一の装置とすることもできる。たとえば、ホスト１００がバーチャリゼーションマネージャを実行しても良い。

【０１４７】

図１４は、本発明を適用した計算機システムの第八の実施形態について示す図である。本実施形態は第七の実施形態と類似した仮想ディスクアレイ装置１５を用いるが、ジャーナル取得・リストア及びジャーナルコピーは、バーチャリゼーションサーバ３００ではなく各ディスクアレイ装置２００上のプログラムによって実行される点が先の実施形態とは異なる。

【０１４８】

又、プライマリサイト１０とセカンダリサイト２０の各ディスクアレイ装置２００は、ファイバチャネルを用いて接続され、ストレージエリアネットワーク（以下ＳＡＮ）を構成する。すなわち、本実施形態において通信リンク６８はＳＡＮである。

【０１４９】

更に、本実施形態では、プライマリサイト１０が有するディスクアレイ装置２００は、自身が有するボリュームが、セカンダリサイト２０のどのディスクアレイ装置２００に対応するか、すなわち通信相手となるディスクアレイ装置２００がどれかの情報を持たなければならない。またその逆も必要である。そのため、バーチャリゼーションサーバ３００は、お互いに各サイトのアドレスマッピング情報を共有し（ステップ３０００）、アドレスマッピング情報の変更も共有する。このマッピング情報は各ディスクアレイ装置２００に通知される。

【０１５０】

本実施形態は第七の実施形態と比較して、ジャーナルバックアップ・リストア処理を各ディスクアレイ装置２００が行なうので、バーチャリゼーションサーバ３００の負荷が低いこと、またプライマリサイト１０とセカンダリサイト２０間のデータ転送がＳＡＮによって行われるので転送速度が速いことが利点である。

【０１５１】

図１６は、第七及び第八の実施形態における、ホスト１００と、そのホスト１

００に対応づけられる仮想ストレージイメージを構成する各ディスクアレイ装置２００のアドレスマッピングを示すテーブル１７０を示す図である。図１３や図１４にはセカンダリホストは１００Ｂの１つしか示していないが、本テーブルでは二つのセカンダリホスト１００Ｂ、１００Ｃに仮想ストレージイメージを提供する場合を示している。

【０１５２】

テーブル１７０には、（１）仮想ストレージイメージを提供する対象のホスト１００に関する列１７２、（２）ホスト１００がアクセスする論理ボリューム（以下「ホストアクセスＬＵ」）に関する列１７４、（３）仮想ストレージイメージを構成するディスクアレイ装置２００に関する列１７６、（４）各ディスクアレイ装置２００上の論理ボリューム（以下「ストレージ装置ＬＵ」）に関する列１７８がある。

【０１５３】

尚、ホスト１００とディスクアレイ装置２００のアドレスの対応については、図１７０に示したようなテーブルでなくとも、同様の情報を持つデータ構造、たとえばポインタによるリストを保持することで実現しても良い。

【０１５４】

上述の実施形態では、形成コピーはジャーナル処理とは別の処理である。しかし、PVOL２２１２の初期状態を示すジャーナル（以下「基底ジャーナル」）を作成して、ペア生成の後生成された更新ジャーナル（update journal）と共にジャーナル処理することでSVOL２２１４へPVOL２２１２の初期データを反映し、形成コピーをジャーナル処理内に包含しても良い。更新ジャーナルは、ペア生成の後、ホストから発行されたライトコマンドに対応するジャーナルである。本発明の説明では、更新ジャーナルは「ジャーナル」または「更新ジャーナル」と呼ぶ。しかし、基底ジャーナルは「基底ジャーナル」としか呼ばない。この用語の区別は、発明の実施の形態でのみ適用し、請求項には適用しない。従って請求項で用いられる「ジャーナル」は基底ジャーナル、更新ジャーナル、マーカジャーナル（後述）、またはそれらの組合わせを示す。

【０１５５】

具体的には、プライマリホスト１００Ａが、プライマリディスクアレイ装置２００Ａに対し基底ジャーナル生成コマンドを発行する。これを受信したプライマリディスクアレイ装置２００Aは、PVOL２１１２の全イニシャルデータから基底ジャーナルを生成する。イニシャルデータは、ペア生成前からPVOL２２１２に存在したデータである。具体的には、基底ジャーナル生成コマンドを受信したプライマリディスクアレイ装置２００Ａは、PVOL２２１２に格納された全てのイニシャルデータを、１つあるいは複数のジャーナルとして正ジャーナルボリュームのジャーナルデータ領域にコピーし、対応するメタデータを各基底ジャーナルに対するメタデータ領域に格納する。ここで、メタデータに含まれるタイムスタンプは、ジャーナルデータがジャーナルデータ領域にコピーされ基底ジャーナルが作成された時刻とする。基底ジャーナルのメタデータにおけるそれ以外の情報（アドレス情報やデータ長など）は更新ジャーナルと同じである。

【０１５６】

基底ジャーナル作成終了後、プライマリディスクアレイ装置２００Ａはプライマリホスト１００Ａに基底ジャーナル生成コマンドの完了を応答する。基底ジャーナルの転送やリストアは、前述の更新ジャーナルに対する処理と同様に行なって良い。

【０１５７】

尚、基底ジャーナルの処理を複数の段階に分け、一回の処理で、イニシャルデータの一部に対して基底ジャーナル作成、ジャーナルコピー、リストアが実行されても良い。例えば正副ジャーナルボリュームの容量がPVOL２２１２やSVOL２２１４よりも小さい場合には、まずPVOLに格納されたデータの前半部分の基底ジャーナルが生成され正ジャーナルボリュームに格納されても良い。その後これらの基底ジャーナル（「第一の基底ジャーナル」）が副ジャーナルボリューム２２２２Bへ転送され、リストアされる。第一の基底ジャーナルの副ジャーナルボリューム２２２２Bへの転送が終了したら、PVOL２２１２に格納されたデータの後半部分の基底ジャーナルが生成され、処理される。

【０１５８】

上述した形成コピーでは、形成コピーとジャーナルリストアが同時に行なわれ

る場合、SVOL2214の各領域のデータが形成コピーのデータでリストアされるかジャーナルボリュームに格納されたデータでリストアされるかを排他的に管理しなければならないが、形成コピーの代わりに基底ジャーナルが生成され処理される場合はそうした管理の必要が無い。

【０１５９】

また、ジャーナルには、基底ジャーナルと更新ジャーナルに加えてマーカジャーナル（marker journal）が含まれても良い。マーカジャーナルは、特殊なジャーナルで、制御情報をジャーナルコピー処理によってプライマリディスクアレイ装置からセカンダリディスクアレイ装置へ伝えるために用いる。マーカジャーナルは、基底ジャーナルや更新ジャーナルや両方のジャーナルと容易に区別できるように、自身のメタデータ内に目的を示すフラグを持つ。マーカジャーナルは予め定められた条件（たとえば基底ジャーナル生成の完了や中断）でプライマリディスクアレイ装置によって生成され、更新ジャーナルも格納される正ジャーナルボリュームへ格納される。

【０１６０】

ジャーナルリストア処理の間、セカンダリストレージサブシステムは、リストアするジャーナルがマーカジャーナルであると判断したならば、それをメモリに格納し、予め定められた処理を実行する（たとえば、マーカジャーナルの内容をセカンダリホスト２００Ｂに報告する）。記憶制御装置２１０がマーカジャーナルを読み、自身のメモリに格納しておいて、ホストから要求を受取ったときにマーカジャーナルの内容をホストへ送信するように実施しても良い。又は、マーカジャーナルの内容のホストへの送信は記憶制御装置２１０が始めても良い。したがって、マーカジャーナルは、セカンダリストレージサブシステムへ、プライマリストレージサブシステムでのデータ処理に関するイベントの情報（たとえば形成コピーや基底ジャーナル生成の完了、形成コピーや基底ジャーナル生成の中断や再開、ジャーナル取得及びその他の処理）を伝える手段となる。

【０１６１】

図１５は、本発明を適用したデータ処理システムの第九の実施形態について示す図である。本実施形態のデータ処理システムは、第一のサイト１０、第二のサ

イト２０及び第三のサイト３０を有する。第一のサイト１０、第二のサイト２０及び第三のサイト３０は、ネットワーク６９（たとえばＳＡＮ）で互いに接続されているとする。本システム５０では、第一のサイト１０と第二のサイト２０から構成される第一のサイトグループ８０１０が定義されているものとする。各サイトはホスト１００を有し、そのホスト１００は非同期コピーマネージャ１５０を有する。第一のサイト１０が有するホスト１００Ａは、プライマリホスト１００Ａとして動作し、第二のサイト２０が有するホスト１００Ｂはセカンダリホスト１００Ｂとして動作する。

【０１６２】

又、其々のサイトのホスト１００は、非同期コピーマネージャ１５０を実行することによって、以下の処理を行う。

【０１６３】

例えば、あるサイトに異常が発生した場合（例えば、セカンダリホスト１００Ｂに接続されたディスクアレイ装置２００Ｂに障害が発生して使用不可能になった場合）、まず、本データ処理システムは、異常が発生したサイトが、プライマリサイト１０か、セカンダリサイト２０かを判断する。本例においては、各サイトの各ホスト１００が自己に接続された機器の異常の発生を監視する構成としても良い。

【０１６４】

異常の発生したサイトがセカンダリサイト２０であれば、異常を検出したホスト１００（ここではセカンダリサイト２０のセカンダリホスト１００Ｂ）が、サイト３０のホスト１００Ｃに対して、プライマリサイト１０と共に新たなサイトグループを形成するよう、要求する（第二のサイトに障害が発生してもデータの複製が行なえるように）。

【０１６５】

異常の発生したサイトがプライマリサイト１０である場合、異常を発見したサイトは、その異常が発生したプライマリサイト１０と組になっているセカンダリサイト２０に対して、プライマリサイトに切り替わるよう、要求を行う。要求を受けたセカンダリサイト２０は、第三のサイトに対して、セカンダリサイトとな

るよう要求する。

【０１６６】

尚、プライマリサイト１０の障害を発見したサイトが、そのプライマリサイト１０と対となるセカンダリサイト２０である場合には、セカンダリサイト２０は、自サイトをプライマリサイトに切り替える。

【０１６７】

本実施形態を利用すれば、例えば、障害時の切り替え用サイトとして、世界各地のデータセンタに、本発明に対応することが可能なサイトを構築し、障害復旧までのレンタルサイトとして顧客に貸し出すことが出来る。このサービスに参加する顧客は、顧客が有するサイトのバックアップ用に、さらには、バックアップサイトを使用した際のさらなるバックアップサイトの確保のために、レンタルサイトを使用することが出来る。また、データセンタを運営するサービスプロバイダは、顧客に対して、レンタルサイトが実際に使用されたことに応じて課金することが出来る。サービスプロバイダは、顧客に対するコンフィギュレーション、たとえばレンタルサイトとバックアップサイトの距離、バックアップサイトの数、バックアップサイトの容量や可用性などに応じて、顧客へ課金を行なっても良い。

【０１６８】

以上、本発明者によってなされた発明を実施例の形態に基づき具体的に説明したが、本発明は前記実施の形態に限定されるものでなく、その要旨を逸脱しない範囲で種々変更可能であることはいうまでもない。

【０１６９】

上述した本発明のデータ処理システムでは、ディスクアレイ装置がジャーナル取得・リストア及びジャーナルコピー処理を行い、ホスト側でジャーナル管理、リモートコピー状態管理を行わせる構成とする。これにより、プライマリサイトとセカンダリサイト間でのデータの複製は、ホスト間で制御命令をやり取りして行われ、実際のデータ転送はディスクアレイ装置間のファイバケーブル等で実施される。このことによって、ホスト間の一般回線のトラフィックを最小限に抑えることができ、またコピーも専用線のため、より高速な回線となり処理性能の向

上が可能である。

【０１７０】

さらに、プライマリサイトとセカンダリサイトのデータ移動には、専用線の他、テープ等の外部記憶装置を使用することで、ユーザ指定の任意の時点のジャーナルを外部記憶装置から読み出してジャーナルリストア処理に用いることが可能になる。この機能により、ユーザの必要な時点のデータへの回復も可能とする効果がある。

【０１７１】

さらに、ディスクアレイ装置がジャーナルをライトコマンドにより他のディスクアレイ装置に書き出す機能を有し、そのデータをホストで読み込み、リストアすることで、セカンダリサイトのディスクアレイに特別な機能を持たせなくてもデータ転送及びデータの複製を実現することが可能となる。

【０１７２】

【発明の効果】

本発明によって、複数サイト間でのデータ転送又はデータ複製をする際に、ホスト間の一般回線のトラフィックを抑えることができ、またデータ転送性能が向上する。さらに、ユーザの必要な時点のデータの回復もできる。

【０１７３】

また、多種多様なサイト間でのデータの複製を容易に実行することが出来る。

【０１７４】

さらに、セカンダリサイトのディスクアレイ装置に特別な機能を持たせなくても良い。つまり、本発明によれば、一般的に互換性の無い例えば製造元の異なるディスクアレイ装置間の接続することが容易に行なえる。

【図面の簡単な説明】

【図１】

本発明を適用したデータ処理システムの第一の実施形態のハードウェア構成を示す図である。

【図２】

図１に示した本発明の第一の実施形態における処理手順を示したフローチャー

トである。

【図３】

図１に示した本発明の第一の実施形態におけるジャーナルの取得、コピー及びリストアの手順を示す図である。

【図４】

図1に示した本発明の第一の実施形態におけるPVOLと正ジャーナルボリューム２２２２Ａの対応関係を示す図である。

【図５】

本発明における正副ジャーナルのジャーナルデータ領域の対応関係を示した図である。

【図６】

本発明の第一の実施形態におけるジャーナル取得、データ転送、ジャーナルリストアの手順の詳細を示すフローチャートである。

【図７】

本発明を適用したデータ処理システムの第二の実施形態を示す図である。

【図８】

本発明を適用したデータ処理システムの第三の実施形態を示す図である。

【図９】

本発明を適用したデータ処理システムの第四の実施形態を示す図である。

【図１０】

本発明を適用したデータ処理システムの第一の実施形態の論理構成を示す図である。

【図１１】

本発明を適用したデータ処理システムの第五の実施形態を示す図である。

【図１２】

本発明を適用したデータ処理システムの第六の実施形態を示す図である。

【図１３】

本発明を適用したデータ処理システムの第七の実施形態を示す図である。

【図１４】

本発明を適用したデータ処理システムの第八の実施形態を示す図である。

【図１５】

本発明を適用したデータ処理システムの第九の実施形態を示す図である。

【図１６】

バーチャリゼーションサーバ３００Ｂ内のアドレステーブルを示す図である。

【図１７】

本発明における副ジャーナルボリュームとSVOLの対応関係を示す図である。

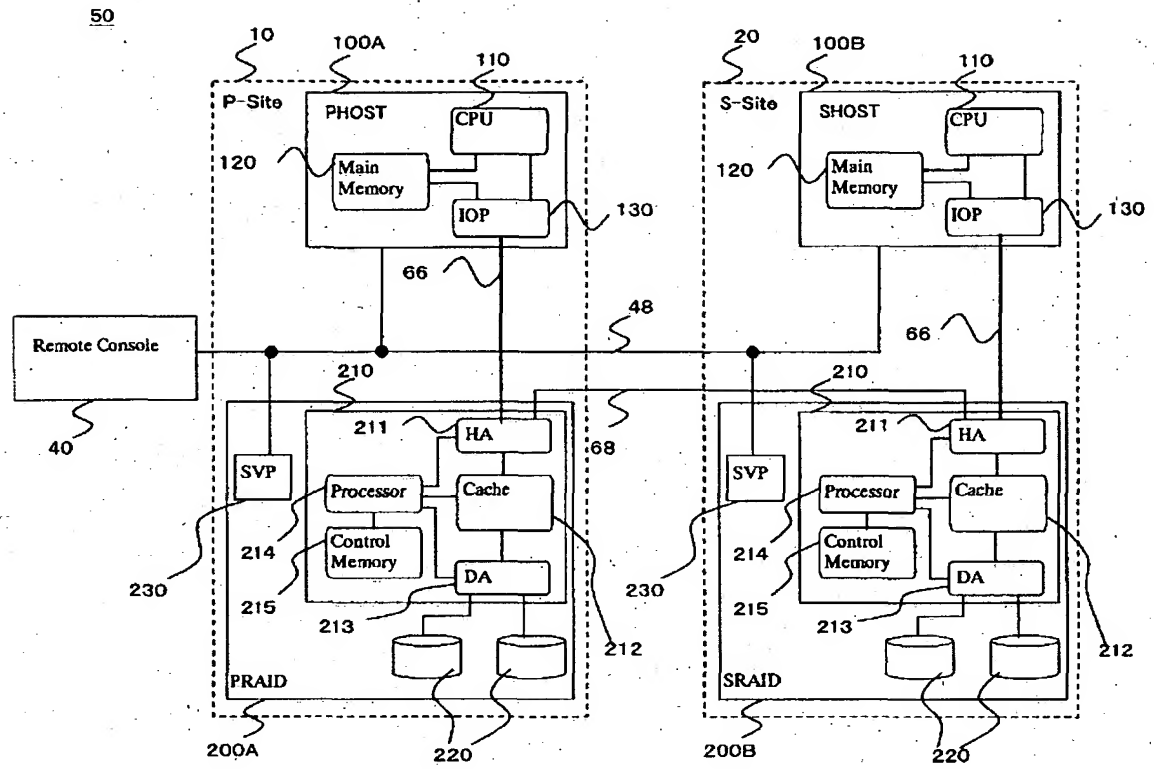
【符号の説明】

１０・・・プライマリサイト、２０…セカンダリサイト、１００Ａ…プライマリホスト、２００Ａ…プライマリディスクアレイ装置、１００Ｂ…セカンダリホスト、２００Ｂ…セカンダリディスクアレイ装置。

【書類名】　図面

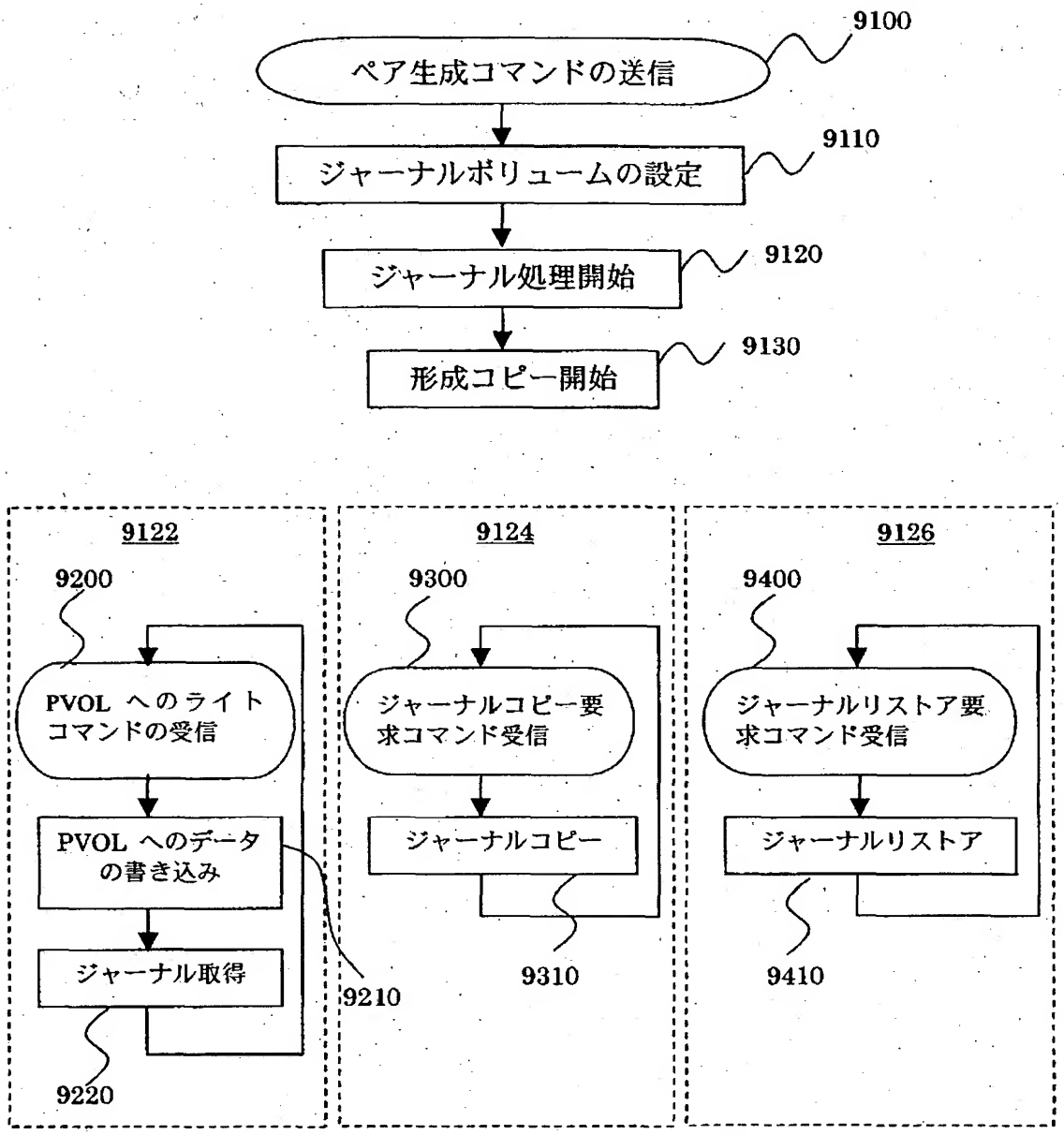
【図１】

図　1


50
10
100A
110
P-Site
PHOST
CPU
120
Main Memory
IOP
130
66
48
Remote Console
40
210
211
HA
SVP
Processor
Cache
214
Control Memory
230
215
DA
213
212
PRAID
200A
220
20
100B
110
S-Site
SHOST
CPU
120
Main Memory
IOP
130
66
210
211
HA
68
SVP
Processor
Cache
214
Control Memory
230
215
DA
213
212
SRAID
200B
220

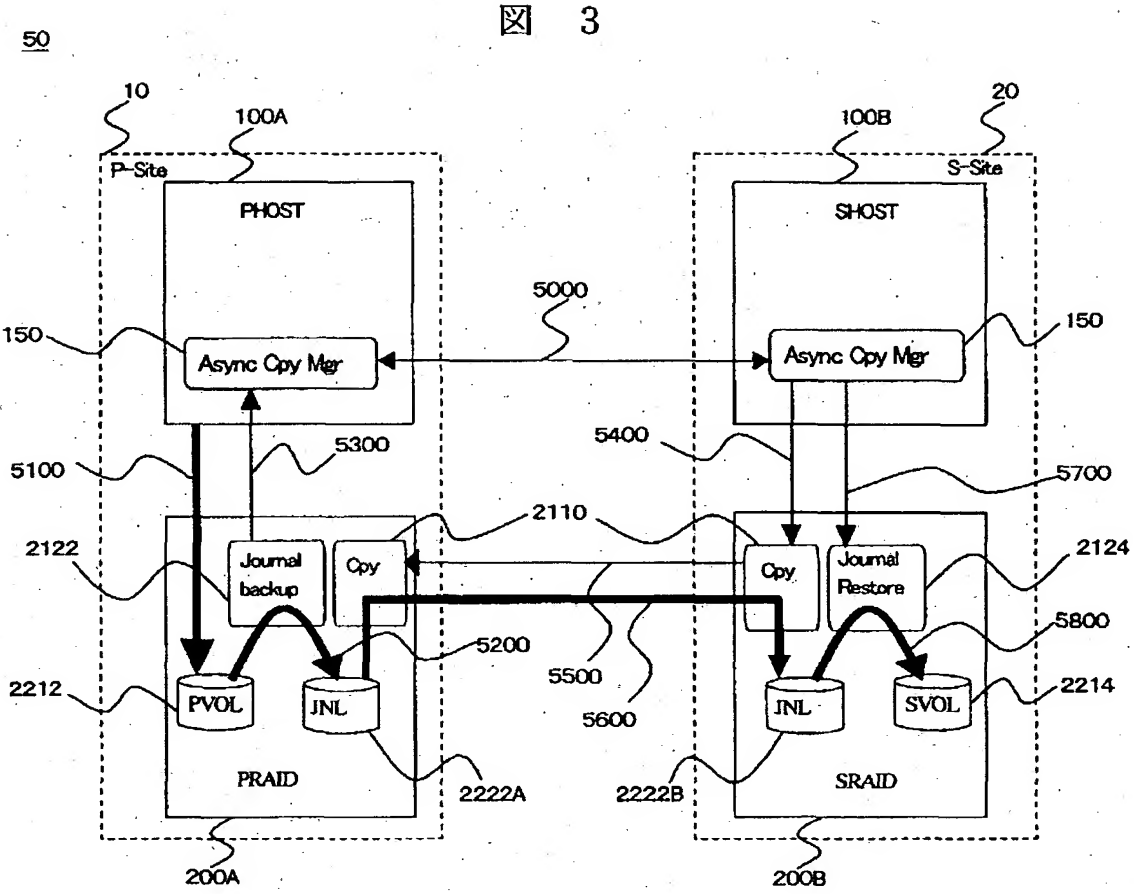
【図２】

図　２


9100
ペア生成コマンドの送信
9110
ジャーナルボリュームの設定
9120
ジャーナル処理開始
9130
形成コピー開始
9122
9200
PVOL へのライトコマンドの受信
PVOL へのデータの書き込み
ジャーナル取得
9210
9220
9124
9300
ジャーナルコピー要求コマンド受信
ジャーナルコピー
9310
9126
9400
ジャーナルリストア要求コマンド受信
ジャーナルリストア
9410

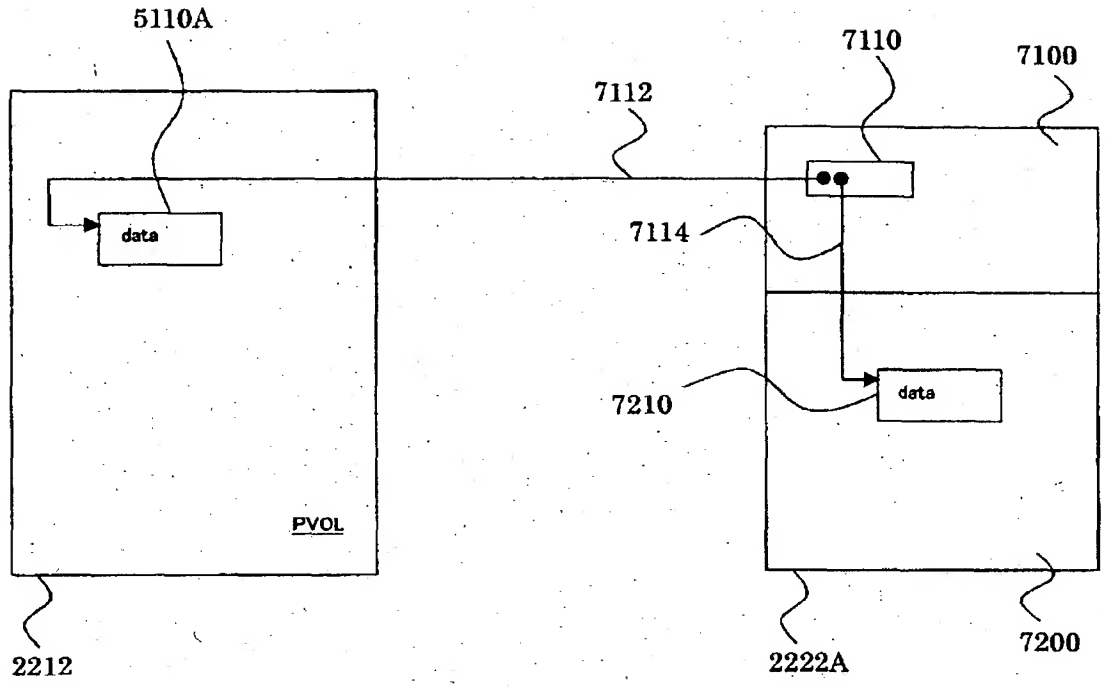
【図３】

図　3


50
10
100A
P-Site
PHOST
150
Async Copy Mgr
5000
100B
20
S-Site
SHOST
150
Async Copy Mgr
5100
5300
5400
5700
2110
2122
Journal backup
Copy
Copy
Journal Restore
2124
5800
5200
5500
5600
2212
PVOL
JNL
JNL
SVOL
2214
PRAID
2222A
2222B
SRAID
200A
200B

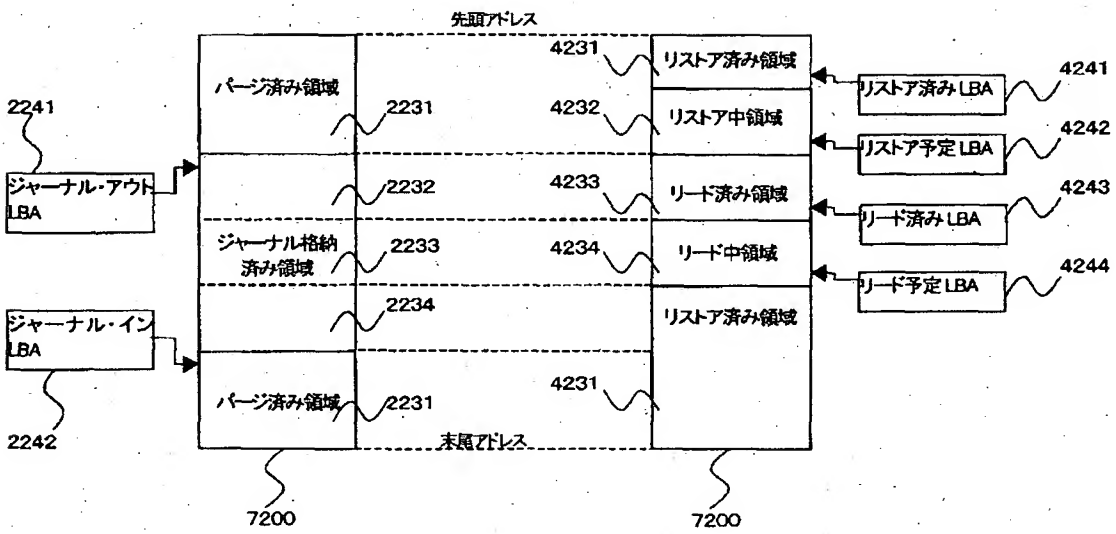
【図４】

図　４


5110A
7110
7100
7112
data
7114
7210
data
PVOL
2212
2222A
7200

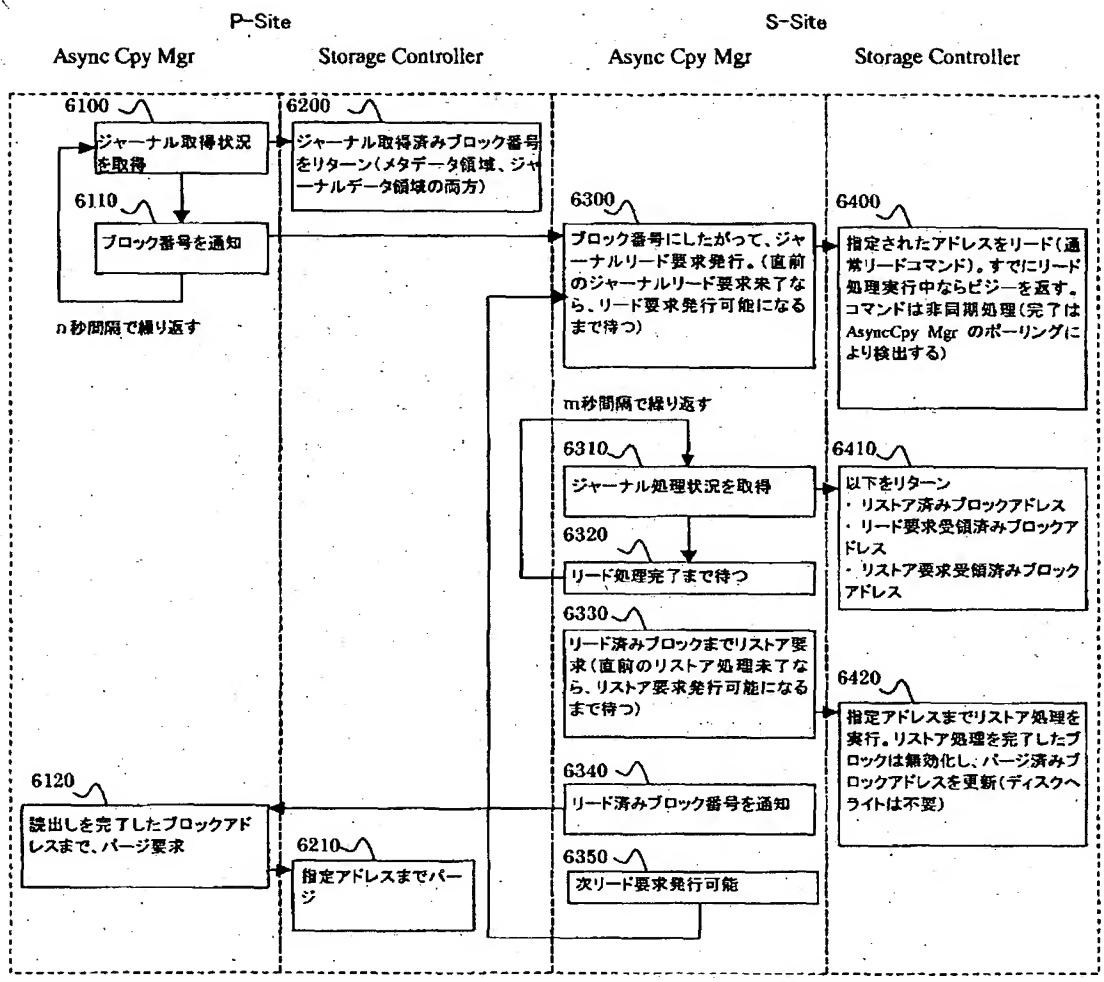
【図５】

図　５


先頭アドレス
4231
リストア済み領域
4241
リストア済みLBA
パージ済み領域
2231
4232
リストア中領域
4242
リストア予定LBA
2241
ジャーナル・アウトLBA
2232
4233
リード済み領域
4243
リード済みLBA
ジャーナル格納済み領域
2233
4234
リード中領域
4244
リード予定LBA
2234
リストア済み領域
ジャーナル・インLBA
パージ済み領域
2231
4231
末尾アドレス
2242
7200
7200

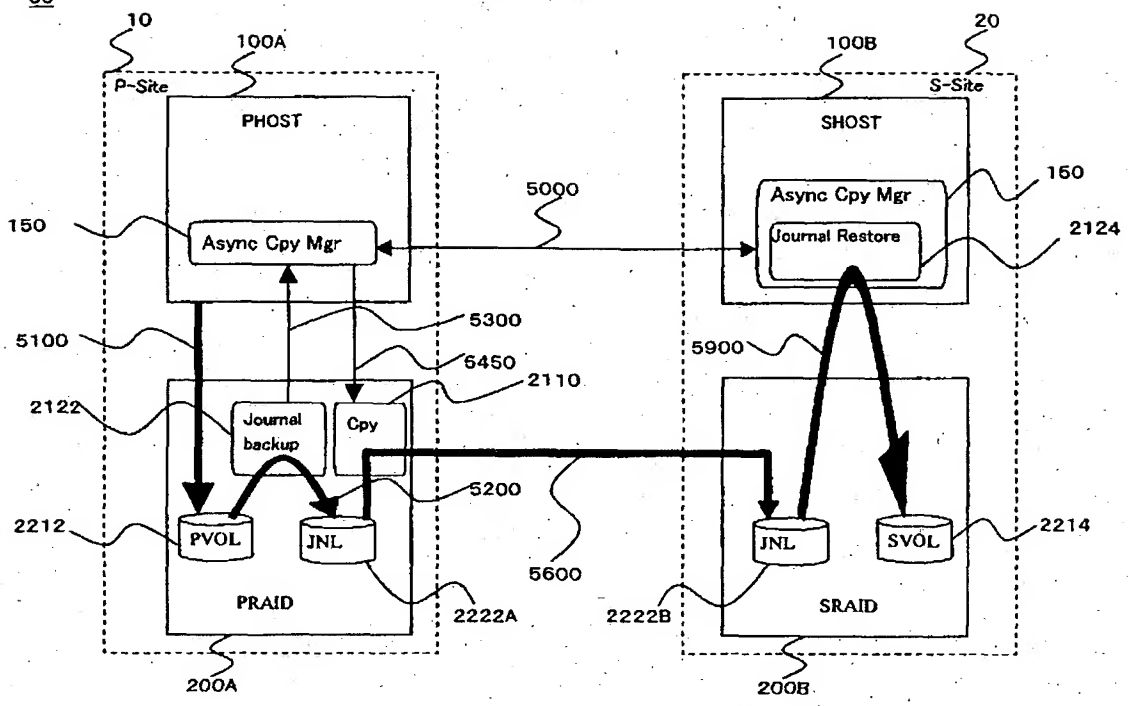
【図６】

図　６


P-Site
S-Site
Async Cpy Mgr
Storage Controller
Async Cpy Mgr
Storage Controller
6100
ジャーナル取得状況を取得
6110
ブロック番号を通知
n秒間隔で繰り返す
6200
ジャーナル取得済みブロック番号をリターン（メタデータ領域、ジャーナルデータ領域の両方）
6300
ブロック番号にしたがって、ジャーナルリード要求発行。（直前のジャーナルリード要求未了なら、リード要求発行可能になるまで待つ）
6400
指定されたアドレスをリード（通常リードコマンド）。すでにリード処理実行中ならビジーを返す。コマンドは非同期処理（完了はAsyncCpy Mgr のポーリングにより検出する）
m秒間隔で繰り返す
6310
ジャーナル処理状況を取得
6320
リード処理完了まで待つ
6410
以下をリターン
・リストア済みブロックアドレス
・リード要求受領済みブロックアドレス
・リストア要求受領済みブロックアドレス
6330
リード済みブロックまでリストア要求（直前のリストア処理未了なら、リストア要求発行可能になるまで待つ）
6420
指定アドレスまでリストア処理を実行。リストア処理を完了したブロックは無効化し、パージ済みブロックアドレスを更新（ディスクへライトは不要）
6120
読出しを完了したブロックアドレスまで、パージ要求
6340
リード済みブロック番号を通知
6210
指定アドレスまでパージ
6350
次リード要求発行可能

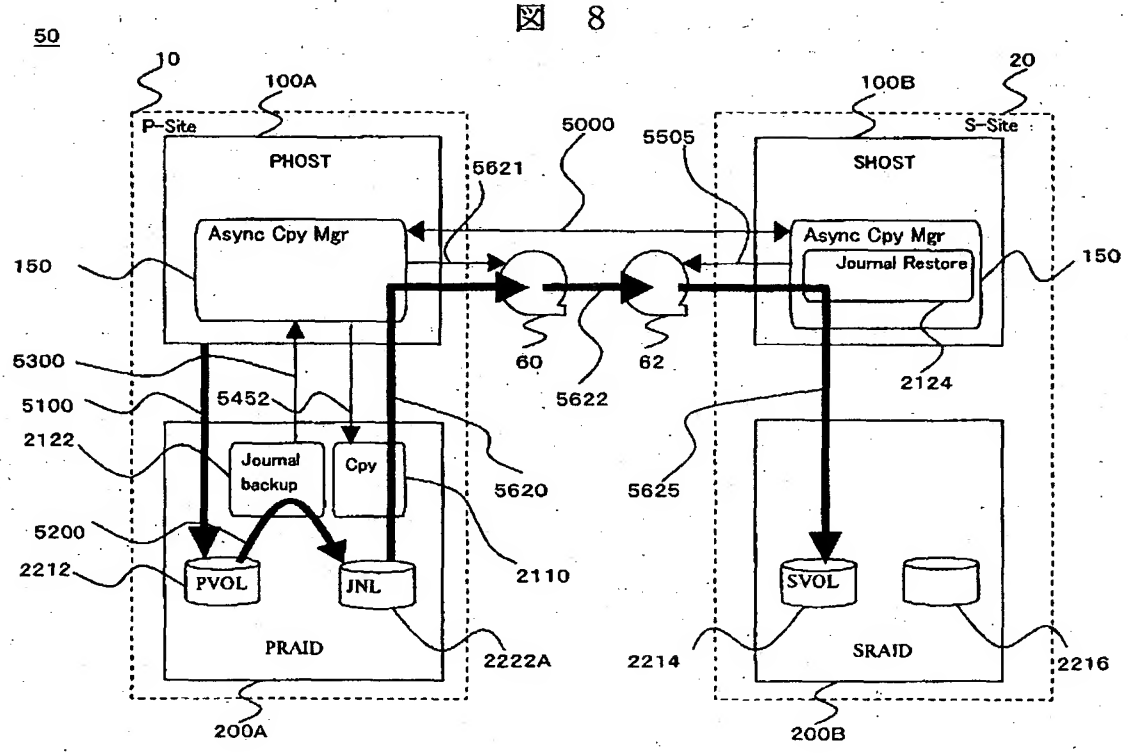
【図７】

図　7


50
10
100A
P-Site
PHOST
150
Async Cpy Mgr
5000
100B
20
S-Site
SHOST
150
Async Cpy Mgr
Journal Restore
2124
5100
5300
5450
2110
5900
2122
Journal backup
Cpy
5200
2212
PVOL
JNL
2214
5600
JNL
SVOL
PRAID
2222A
2222B
SRAID
200A
200B

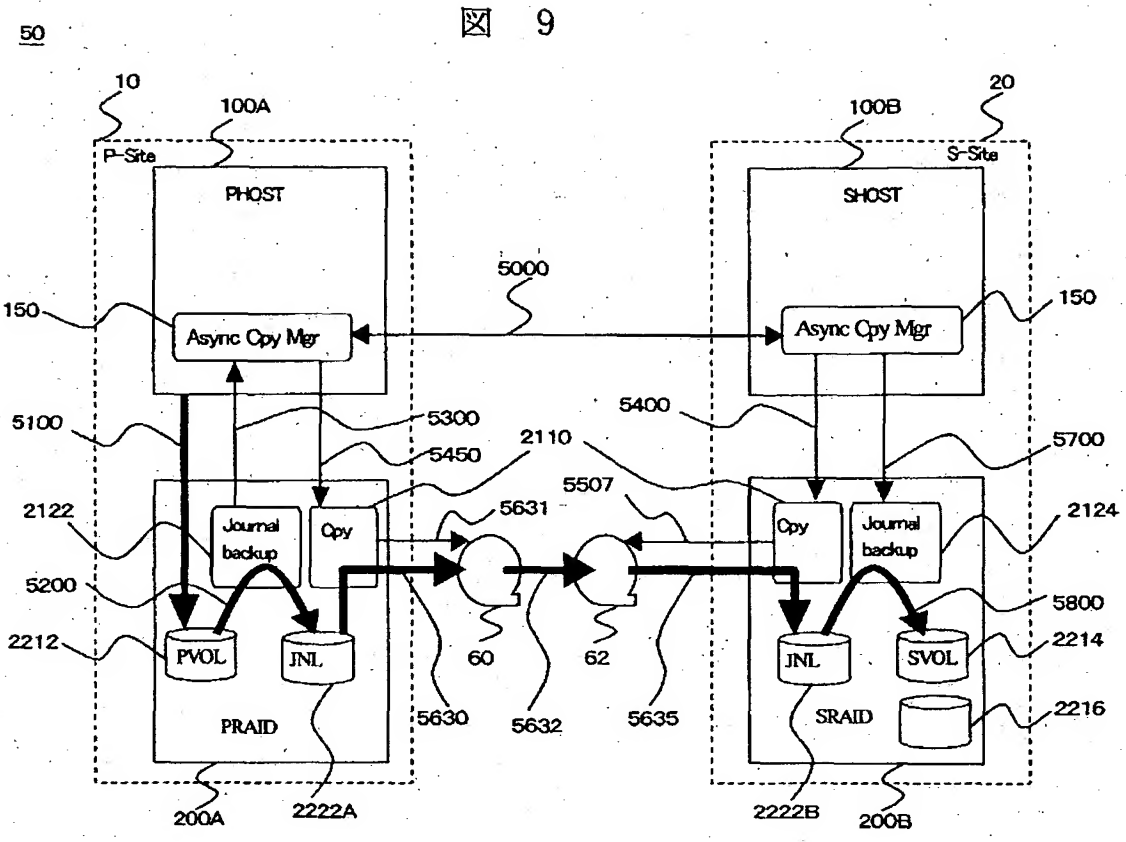
【図８】

図　8


50
10
100A
P-Site
PHOST
5000
5621
5505
100B
20
S-Site
SHOST
Async Cpy Mgr
150
Async Cpy Mgr
Journal Restore
150
60
62
5622
2124
5300
5452
5100
2122
Journal backup
Cpy
5620
5625
5200
2212
PVOL
JNL
2110
SVOL
PRAID
2222A
2214
SRAID
2216
200A
200B

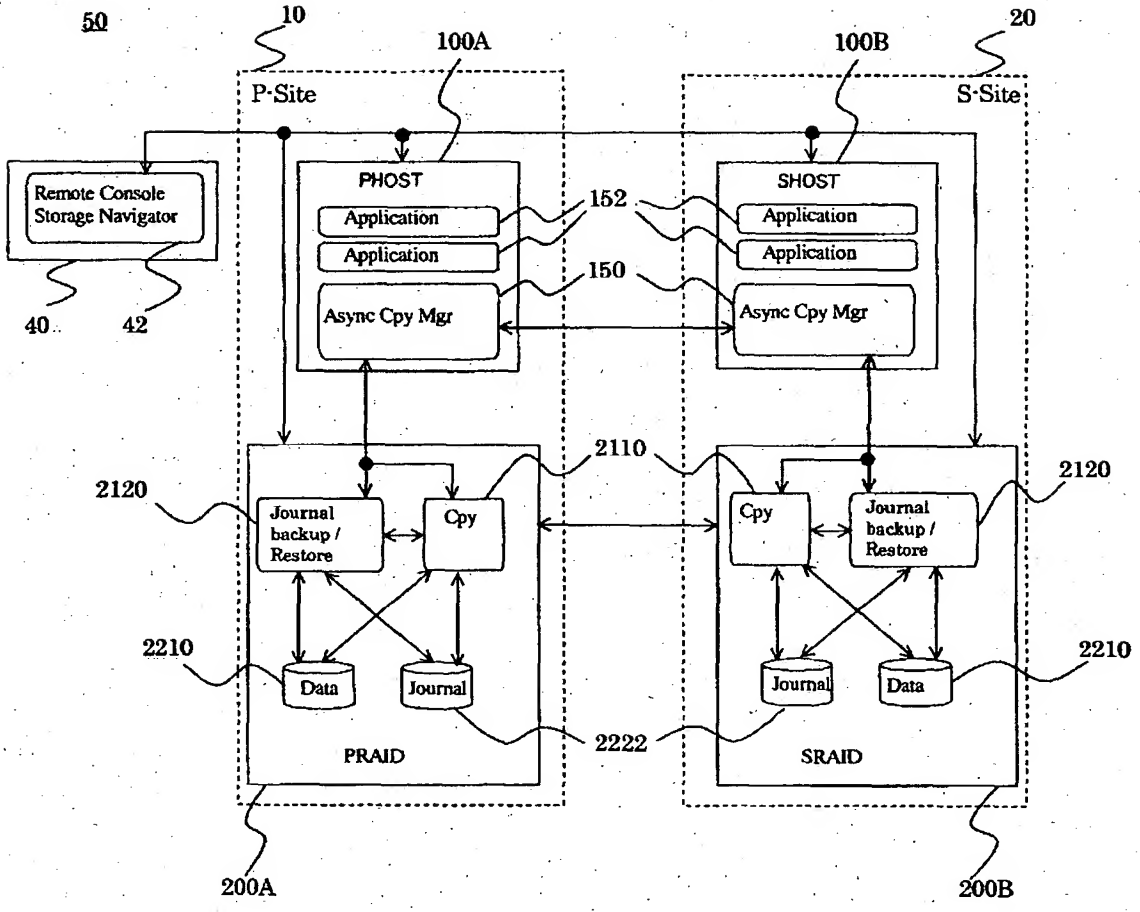
【図９】

図　9


50
10
100A
P-Site
PHOST
150
Async Cpy Mgr
5000
100B
20
S-Site
SHOST
150
Async Cpy Mgr
5100
5300
5450
2110
5400
5700
5507
2122
Journal backup
Cpy
5631
Cpy
Journal backup
2124
5200
5800
2212
PVOL
JNL
60
62
JNL
SVOL
2214
PRAID
5630
5632
5635
SRAID
2216
200A
2222A
2222B
200B

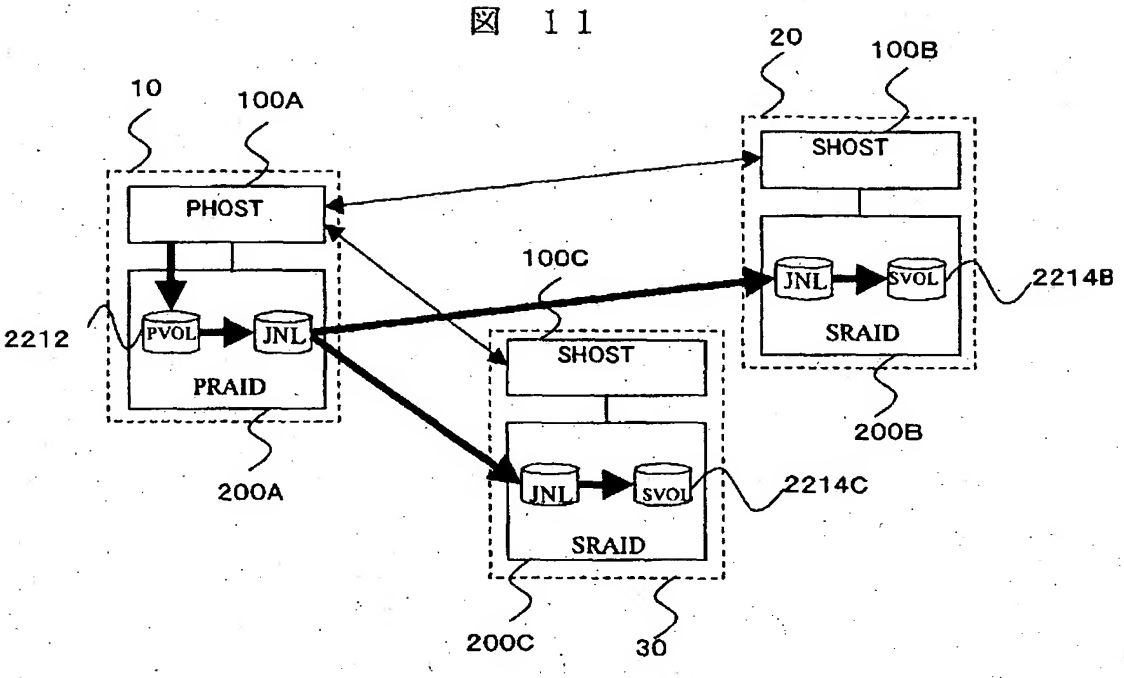
【図１０】

図　１０


50
10
100A
P-Site
100B
20
S-Site
Remote Console
Storage Navigator
40
42
PHOST
Application
Application
Async Cpy Mgr
152
150
SHOST
Application
Application
Async Cpy Mgr
2110
2120
Journal
backup /
Restore
Cpy
Cpy
Journal
backup /
Restore
2120
2210
Data
Journal
2222
Journal
Data
2210
PRAID
SRAID
200A
200B

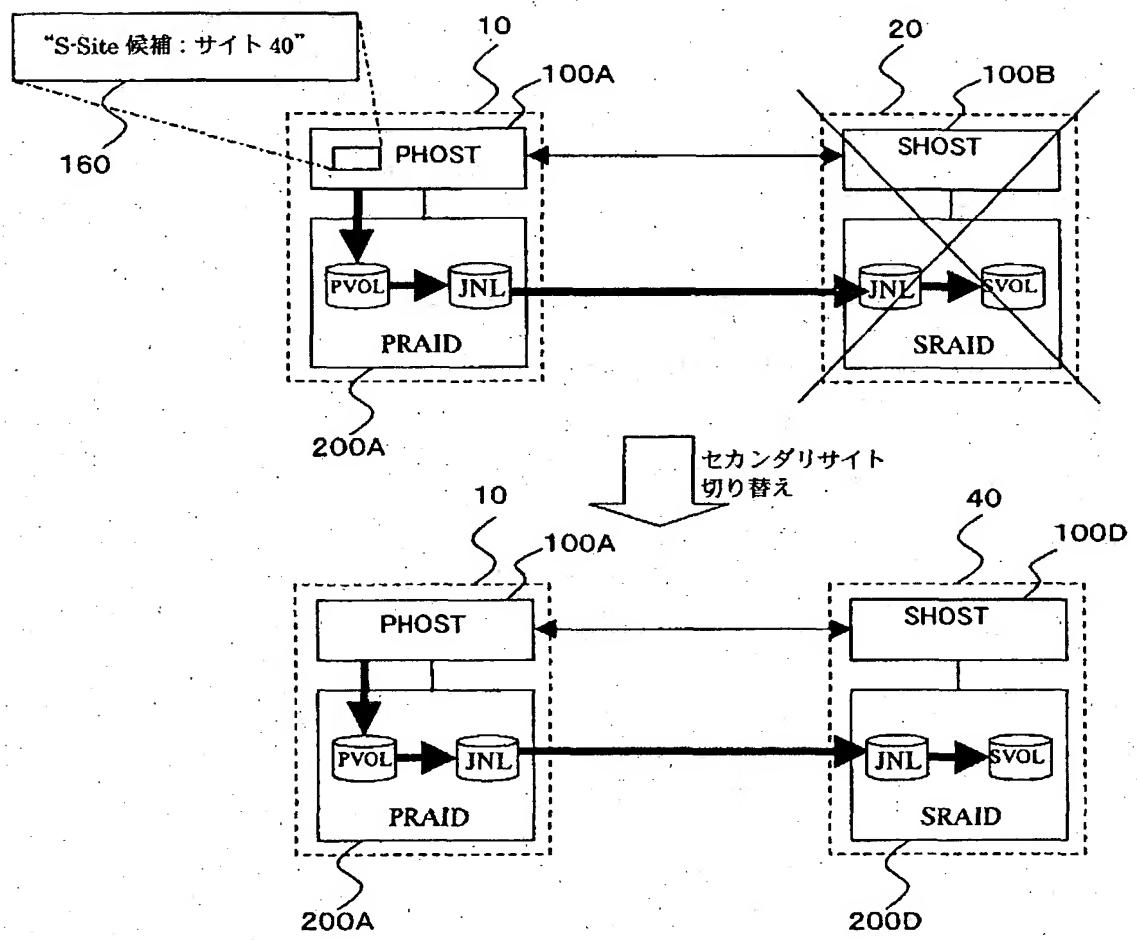
【図１１】

図　１１


10
100A
PHOST
2212
PVOL
JNL
PRAID
200A
20
100B
SHOST
JNL
SVOL
2214B
SRAID
200B
100C
SHOST
JNL
SVOL
2214C
SRAID
200C
30

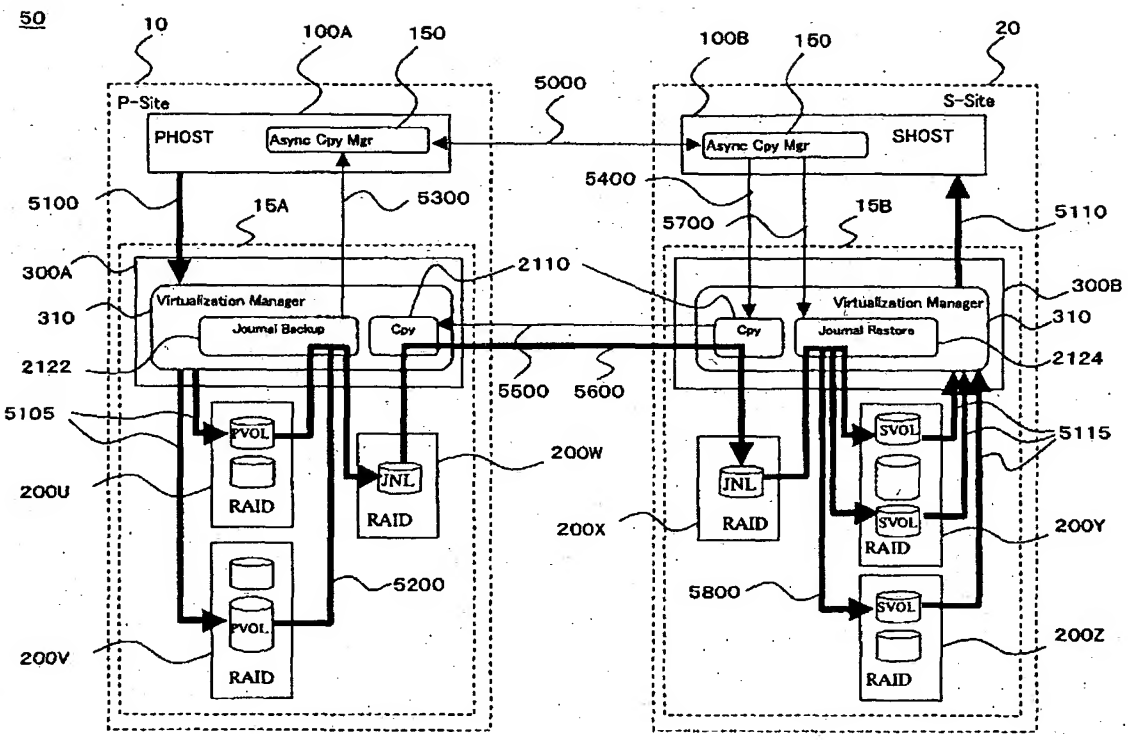
【図１２】

図　１２


"S-Site 候補：サイト 40"
160
10
100A
20
100B
PHOST
SHOST
PVOL
JNL
JNL
SVOL
PRAID
SRAID
200A
セカンダリサイト
切り替え
10
100A
40
100D
PHOST
SHOST
PVOL
JNL
JNL
SVOL
PRAID
SRAID
200A
200D

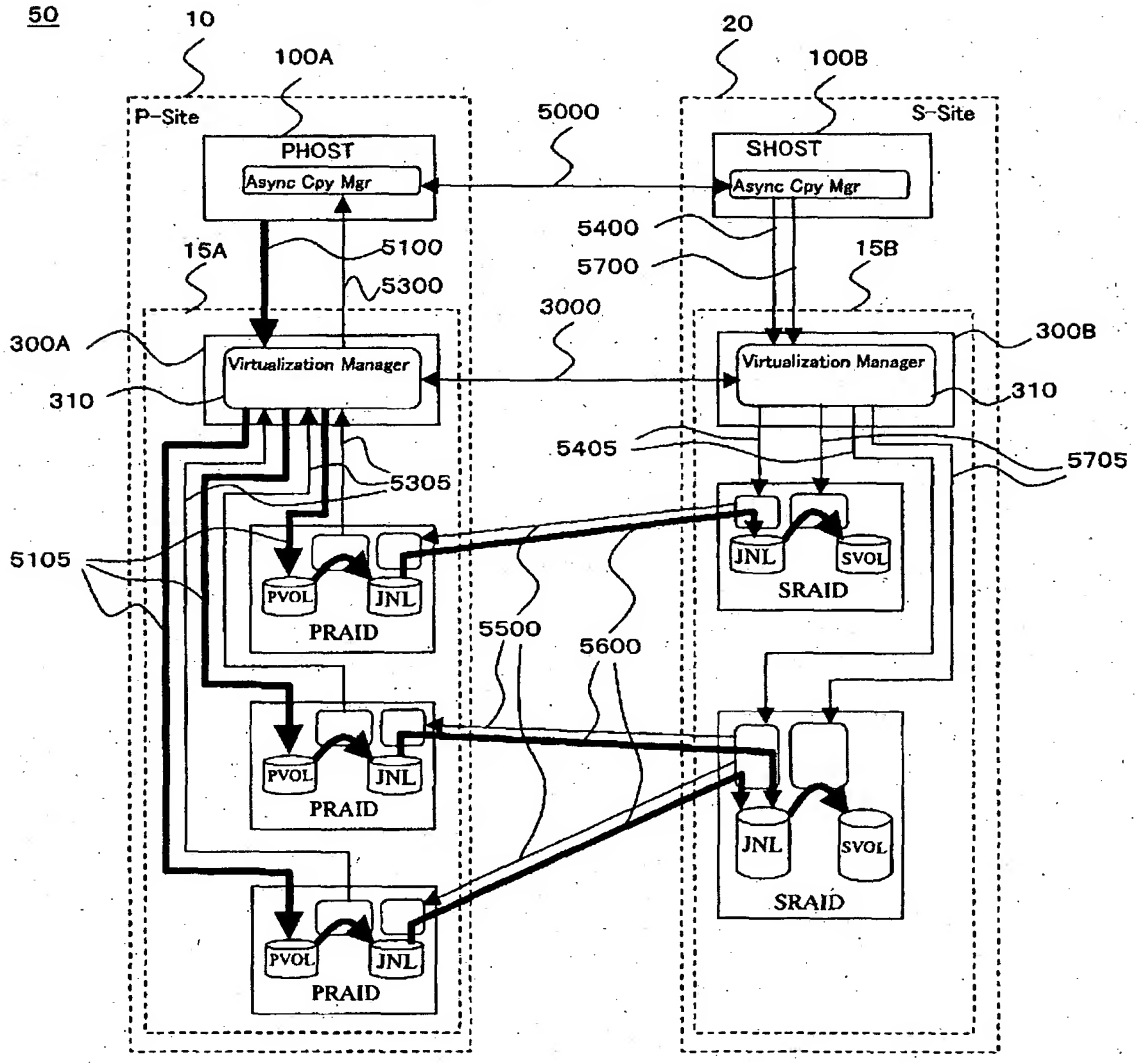
【図１３】

図　１３


50
10
100A
150
100B
150
20
5000
P-Site
S-Site
PHOST
Async Cpy Mgr
Async Cpy Mgr
SHOST
5100
15A
5300
5400
5700
15B
5110
300A
2110
300B
310
Virtualization Manager
Journal Backup
Cpy
Cpy
Journal Restore
Virtualization Manager
310
2122
2124
5500
5600
5105
5115
PVOL
SVOL
200W
200U
JNL
JNL
RAID
RAID
RAID
RAID
200X
200Y
5200
5800
PVOL
SVOL
200V
200Z
RAID
RAID

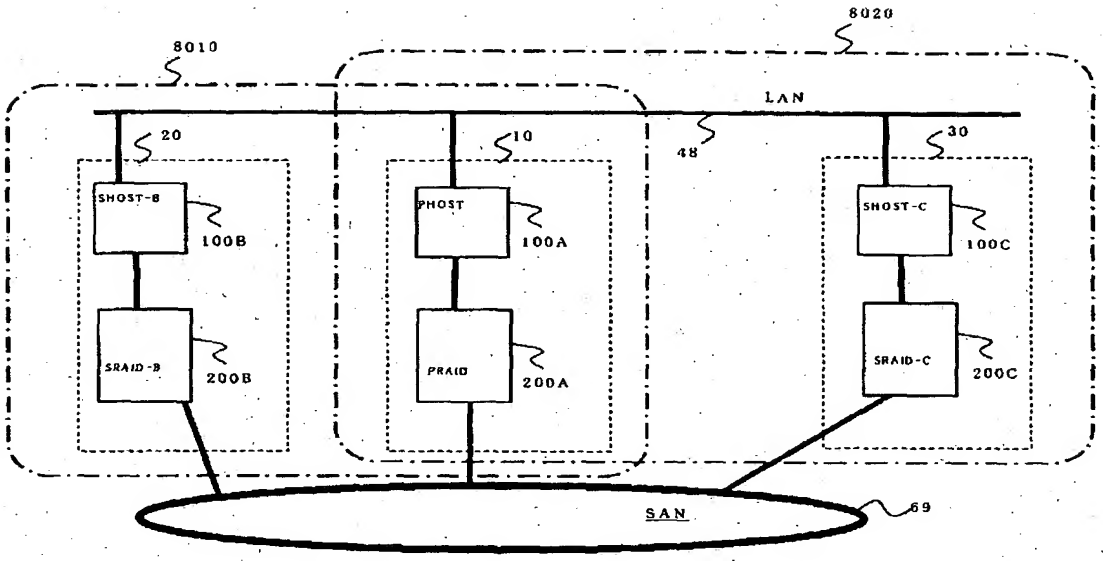
【図１４】

図　１４


50
10
100A
P-Site
PHOST
Async Cpy Mgr
15A
5100
5300
300A
Virtualization Manager
310
5305
5105
PVOL
JNL
PRAID
PRAID
PRAID
5000
20
100B
S-Site
SHOST
Async Cpy Mgr
5400
5700
15B
3000
300B
Virtualization Manager
310
5405
5705
JNL
SVOL
SRAID
JNL
SVOL
SRAID
5500
5600

【図１５】

図　１５



【図１６】

図　１６

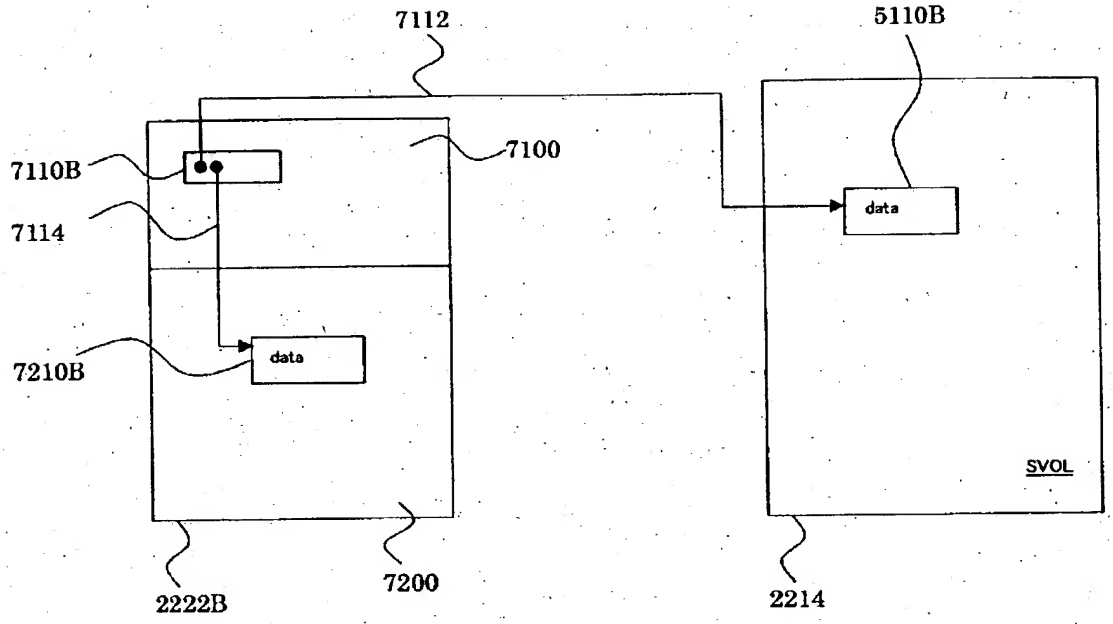
Host	Host access LU	Storage device	Storage device LU
１００Ｂ	０～３	２００Ｘ	０～３
	４～７	２００Ｙ	０～３
	８～１１	２００Ｙ	１２～１５
	１２～１５	２００Ｚ	１６～１９
１００Ｃ	０～７	２００Ｙ	４～１１
	８～１１	２００Ｚ	４～７

170

172　174　176　178

【図１７】

図　１７


7112
5110B
7110B
7100
7114
data
7210B
data
SVOL
2222B
7200
2214

【書類名】　要約書

【要約】

【課題】

ホストベースリモートサイトへのデータの複製は、ホストに負荷がかかる。

【解決手段】

実際のデータ転送は各サイトが有する記憶装置間で、その制御等はホストで行うことにより、ホストに負荷をかけずに、より複雑なシステムにおいても、リモートサイトへデータを複製することが出来る。また、ジャーナルを用いることにより、ユーザの要求に合ったデータの複製を作成することができる。

【選択図】　図１

認定・付加情報

特許出願の番号	特願２００３－０５０２４４
受付番号	５０３００３１３９５２
書類名	特許願
担当官	第七担当上席　　００９６
作成日	平成１５年　２月２８日

＜認定情報・付加情報＞

【提出日】　平成15年 2月27日

次頁無

出願人履歴情報

識別番号　〔０００００５１０８〕

１．変更年月日　１９９０年　８月３１日
［変更理由］　新規登録
住　所　東京都千代田区神田駿河台４丁目６番地
氏　名　株式会社日立製作所